ARL-TR-7301 • MAY 2015

US Army Research Laboratory

The Communications and Networks Collaborative Technology Alliance Publication Network: A Case Study on Graph and Simplicial Complex Analysis

by Terrence J Moore, Robert J Drost, and Ananthram Swami

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-7301 • MAY 2015

US Army Research Laboratory

The Communications and Networks Collaborative Technology Alliance Publication Network: A Case Study on Graph and Simplicial Complex Analysis

by Terrence J Moore, Robert J Drost, and Ananthram Swami
Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD MM YYYY)*
May 2015

2. REPORT TYPE
Final

3. DATES COVERED (From To)
May 2012-December 2014

4. TITLE AND SUBTITLE
The Communications and Networks Collaborative Technology Alliance Publication Network: A Case Study on Graph and Simplicial Complex Analysis

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Terrence J Moore, Robert J Drost, and Ananthram Swami

5d. PROJECT NUMBER
R.0014460.11

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
US Army Research Laboratory
ATTN: RDRL-CIN-T
Adelphi, MD 20783-1138

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-7301

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
<terrence.j.moore.civ@mail mil>

14. ABSTRACT
This report examines the co-authorship network from the publication record of the Communication and Networks Collaborative Technology Alliance (C&N CTA). In addition to several traditional graph properties, we examine various simplicial complex properties of the network where the simplex structure is generated by the group of co-authors on individual papers. In particular, we study aspects of the connectivity of k-simplices and k-facets. We study particular centrality characteristics, revealing inherent properties of co-authorship networks, including facet degrees follow a power law distribution, homology cycles intersect, and minimal non-faces are due to independent clustering. We also study more traditional topological properties of the network, including Q-analysis, f-vectors, and the Euler characteristic.

15. SUBJECT TERMS
collaboration network, simplicial complex, Communication and Networks Collaborative Technology Alliance, C&N CTA

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Terrence J Moore
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	UU	60	**19b. TELEPHONE NUMBER (Include area code)** 301-394-1236

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

List of Tables

1. The Communications and Networks Collaborative Technology Alliance

The Communications and Networks Collaborative Technology Alliance (C&N CTA) was a research consortium of academic, industry, and Government research partners funded by the US Army Research Laboratory (ARL) for the purpose of developing "technologies that enable a fully mobile, fully communicating, agile, situationally aware, and survivable lightweight force with Internetworked Command, Control, Communications, Computers, Intelligence, Surveillance, and Reconnaissance (C4ISR) systems. These wireless network technologies were required to operate with a heterogeneous mixture of individual Soldiers, ground vehicles, airborne platforms, unmanned aerial vehicles, robotics, and unattended ground sensor networks; and operate while on-the-move with a highly mobile network infrastructure, under severe bandwidth and energy constraints, while providing secure, jam-resistant communications in noisy hostile wireless environments."[1] The C&N CTA program started in fiscal year 2002, ended in fiscal year 2009 (FY02–FY09), and produced a total of 960 publications by 518 authors.

The C&N CTA dataset includes certain publication metadata for the 8-year run of the program and was collected in an Excel spreadsheet. This spreadsheet consisted of 16 pages, with each page including a single table corresponding to the FY and the publication type (journal or conference proceeding). Each page includes a list of publications with fields for the paper title, paper authors, those authors' affiliations (not always matched to the different authors in the case of a co-authored work), the primary technical area of the paper, the journal or conference proceeding title, and the publication date (this occasionally included a notation indicating that the submission date was entered instead).

There is "noise" observed in the recordkeeping of the tables. The formats of the tables are inconsistent from year to year, often including some redundancy in the information provided. For example, the FY02 Journal table included affiliation information in both the affiliation field and the author field. The entries for the fields are also often inconsistent. For example, the entry for an author's name might include either a first name or first initial; title entries might include simply the title or the complete citation; and publication date entries might include the month, quarter, or just the FY information. As already stated, the submission date was often substituted for the publication date. Also, the dataset includes a number of duplicate entries (e.g., where an entry exists for when the paper was first submitted and another entry exists for when the same paper appeared). Given the size of this dataset, the obvious misspelling errors and duplicate entries are easily correctable; also it is easy to link entries with only an author's first initial and last name to the full name.

The amended entries include 960 publications (292 journal papers and 668 conference proceedings) by 518 authors. These data were imported into MATLAB with each paper recorded as an element in a structure array, with each structure having the following fields:

- ID_no (a number between 1 and 960),
- ID (a tag based on FY, type, and order appearance of the paper on its table page),
- Title,
- Authors,
- Organizations,
- Area (T1-T4),
- BookTitle (primarily journal or conference title),
- Type ("Journal" or "Conference"),
- Date,
- Abstract (if recorded on the Excel sheet), and
- Author.

In addition, for each paper entry, the Author field above was a substructure array whose length was determined by the number of co-authors and that included the following fields:

- ID_no (a number between 1 and 518),
- ID (a tag based on the author's name),
- Name,

- First (name), and
- Last (name).

From these structure fields, any number of cell arrays or other data structures can be generated using the IDs. This report discusses the links between each paper ID and the authors' IDs on that paper.

1.1 Metrics of the C&N CTA Dataset

The C&N CTA dataset includes publication information for each year of the program. This enables consideration of the data in at least 2 ways. First, each year of the data can be examined separately for a variety of metrics and compared with the other years' metrics. Alternatively, the data can be analyzed in a cumulative manner, so as to illuminate how properties evolved as the dataset grew from the program's first year to its completion. We consider both approaches simultaneously throughout this report.

In determining these metrics, we use the data entered in the publication date field, even in instances where the submission date was used in that field. If the paper was listed on a table for a particular FY and publication type, it is treated as belonging to that year's metrics. The only time this is not true is in the rare instance when a paper was duplicated or listed twice, once with the submission date and again with the publication date. In such cases, we use the information for the latter entry and omit the former.

Likely, the dataset contains other errors. For example, a paper might have multiple entries with duplicate titles, i.e., when the title changed between submission and publication. It is also possible that the dataset includes entries for publications that were submitted but never accepted. This last scenario, however, still represents a collaboration, albeit, in some sense, an unsuccessful one. We simply refer to a particular year's listed publications as being published in that year as the bulk of them were. Some publications undoubtedly appeared after the final FY of the program; these are included in the table for the final FY.

1.1.1 Number of Papers and Authors

Basic metrics for the number of papers and number of authors involved in this CTA are given in Table 1. The entries for the numbers of papers in the table represent the number of papers listed in the corresponding spreadsheet page. No distinction is made here between authors who were principal investigators (PIs) or authors who were students or postdocs, nor between authors from academia, industry, or Government.

Table 1 C&N CTA: Paper and author metrics

Fiscal Year	# Papers	# Journ.	# Conf.	# Papers (cum.)	# Authors	# Authors (cum.)
2002	80	17	63	80	101	101
2003	104	16	88	184	99	153
2004	173	62	111	357	184	258
2005	118	30	88	475	143	293
2006	182	74	108	657	199	375
2007	130	37	93	787	136	424
2008	99	34	65	886	155	482
2009	74	22	52	960	105	518
mean	120	36.5	83.5		140.25	
median	111	32	88		139.5	

Some important properties follow. The mean and median number of papers published yearly is 120 and 111, respectively. The mean and median number of authors participating in at least 1 publication yearly is 140.25 and 139.5, respectively. The numbers display a growth in the number of publications from FY02 to FY04 before a dip in FY05, a return to FY04 numbers in FY06, and then a steady decline until the conclusion of the CTA in FY09. A number of causes may have contributed to this behavior.

One explanation involves the student participants. There was undoubtedly some lag time at the start of the program due to the research time required to produce initial results, accounting for the initial upswing. Mid-program, there was likely a phasing out of many student authors (who graduated and left the program) and, thus, a transitional period during which new student authors were acclimated to the program.[2] This could explain the dip from FY04 to FY05 and the return in FY06. The gradual decline after FY06 could then be explained by the lack of new students participating in the program as it wound down.

An alternative explanation might be found from examining the budget numbers for each FY. Less funding could have led to the departure of PIs, less support for students, or even fewer opportunities for collaboration (e.g., if travel funding was curtailed). Program redirections could also be responsible for the reduction of publishing output in FY05, since a program shift redirects funding and requires the usual initialization time of a scientific investigation, as seen at the start of the program.

Of the more than 500 authors who participated at some level in the creation of the publications in the C&N CTA, only 12 published in every year of the program. In alphabetical order, these authors are Paul D Amer, John S Baras, Georgios B Giannakis, Janardhan R Iyengar, Tao Jiang, John E Kleider, Xiaoli Ma, Anthony J McAuley, Tarek N Saadawi, Randall Stewart, Ananthram Swami, and Lang Tong.

1.1.2 Number of Authors per Publication and Publications per Author

The empirical distribution of the number of authors per publication is depicted in Fig. 1. The mean and median number of authors per paper are 2.7708 and 3, respectively. The vast majority of papers are written by either 2 or 3 authors. An estimate of the fraction of papers with k authors for small collaborations in a large network can be derived as[3]

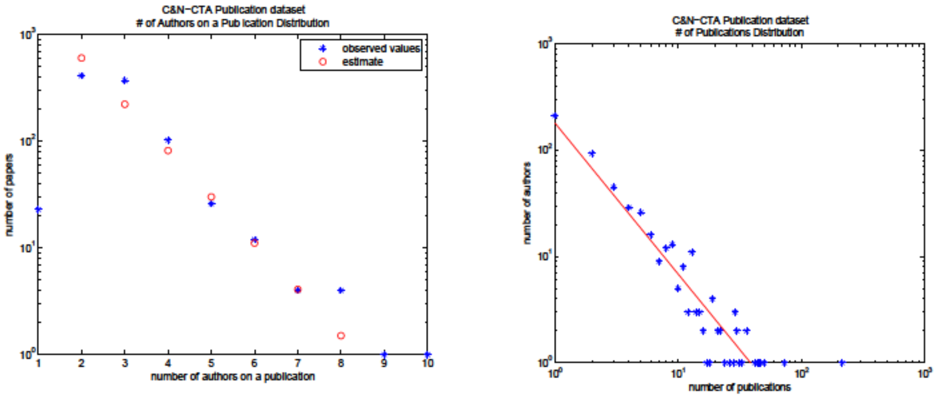
$$p(k) = (e^{\lambda} - 1)e^{-\lambda k}, \tag{1}$$

where λ is the reciprocal of the mean number of authors on a paper. For this dataset, $\lambda = 1/2.7708 = 0.3609$ produces an estimate that decays too slowly, so instead the plot depicts the above function with $\lambda = 1$ (rescaled to the total number of papers instead of the fraction). It should not be inferred that this implies the program led to fewer large collaborative teams than expected, as the C&N CTA publication network has an additional restriction relative to general publication networks, namely, that collaborations are restricted to papers that included participants as co-authors and were relevant to the program.

The empirical distribution of the number of publications an author contributed to is displayed in Fig. 1. The solid line, given by

$$\text{\# authors who published } x \text{ papers} = 180.89x^{-1.42}, \tag{2}$$

is an estimate of the slope from a linear regression excluding the outlier with the largest residual. The mean, median, and mode of the number of publications per author are 5.1351, 2, and 1, respectively. The large number of single-publication authors is primarily due to students who had a limited role in the program. How-



(a) Number of authors per publication (b) Number of Publications Distributions

Fig. 1 The C&N CTA: Number of authors per publication and publications per author

ever, it is also likely that some work was "piggy-backed" on other (non-C&N CTA) research efforts, which would include collaborations with academia, industry, and Government outside of the core C&N CTA members. Excluding authors with only a single publication, which are likely candidates to be either researchers not funded by the program or students, the mean and median number of publications per author jump to 7.9772 and 4, respectively. These numbers compare favorably to studies in other scientific fields,[4] especially considering that the members of the C&N CTA may have had other publications in the field not relevant to the program.

1.1.3 Publication Ranks

In a publications dataset, one measure of key authors is the number of publications they account for in the dataset. If they are influential, they will account for the bulk of the publications. Table 2 provides a list of the top 21 authors by publication rank, i.e., the authors who published the most papers as part of the C&N CTA program. These top 21 authors collectively contributed to 680 or 70.83% of the papers. The top 10 authors collectively contributed to 471, or nearly half (49.06%) of the papers. Amazingly, the top author (Giannakis) co-authored as many papers as the second through fifth most-published authors combined. He contributed to 22.4% of the publications in the dataset.

Table 2 C&N CTA: Number of publications top-21 list

Rank	Author	# of Papers
1	Georgios B Giannakis	215
2	Lang Tong	73
3	Qing Zhao	50
4	Xiaoli Ma	47
5	Ananthram Swami	45
6	Gordon L Stuber	44
7	Shengli Zhou	42
8	Brian M Sadler	36
8	John S Baras	36
10	Zhengyuan Xu	33
11	Myung Jong Lee	31
12	Anthony J McAuley	30
12	Tarek N Saadawi	30
14	Yingbo Hua	29
14	John E Kleider	29
14	Sergio Verdu	29
17	Mariusz A Fecko	28
18	Paul D Amer	26
19	Adarshpal S Sethi	24
20	Sunil Samtani	22
21	Alenka G Zajic	22

2. Graph Analysis of the C&N CTA Publication Network

A graph can be created from the co-authorship dataset by identifying vertices in the graph with authors and identifying edges with the co-authorship relation,[5–8] i.e., an edge exists between 2 vertices if and only if the 2 authors corresponding to the 2 vertices have co-authored at least 1 paper in the dataset.[9] (Naturally, if an author published with no co-authors then the vertex corresponding to that author is isolated.) For the study here, we model the C&N CTA as a simple undirected graph, although it is possible to consider the frequency or totality of an author pair's co-authorship using weighted directed graphs.[10]

This construction on the C&N CTA dataset generates a graph of 518 vertices and 1248 edges; a visualization is depicted in Fig. 2. Naturally, many edges exist for certain papers (i.e., papers with at least 3 co-authors), and many papers might be denoted by a single edge (i.e., authors who co-authored more than once).

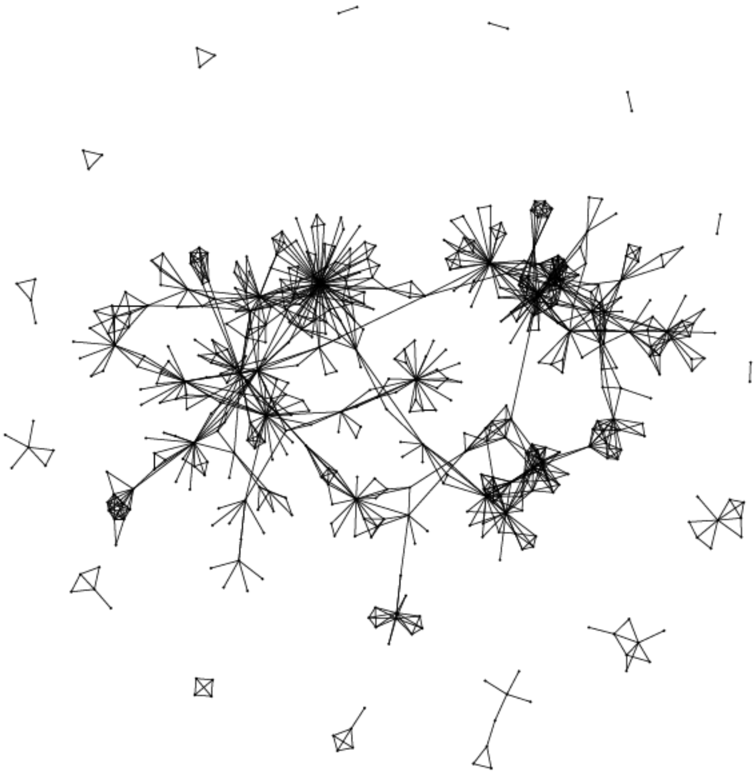


Fig. 2 The C&N CTA graph

We study 2 important types of subgraphs on the C&N CTA network: 1) the subgraph of vertices and edges for a particular FY, and 2) the subgraph of the vertices and edges from the first FY to a particular FY, e.g., FY02 to FY05.

We assume the reader has a general knowledge of much of the graph terminology used in this chapter. If there are any definitions the reader is unfamiliar with, these can be found in other sources.[6–8]

2.1 Connectivity

In this section, we examine the evolution of various aspects of the network, such as the number of vertices and edges, the average path length, the number of components, density, diameter, and clustering.

2.1.1 Network Size

We have already stated that the graph of the C&N CTA dataset for the program duration has 518 vertices and 1248 edges. The evolution of the number of vertices and edges of this network graph over the program lifetime is given in Table 3. The "per year" (or yearly) numbers only include the authors' interactions recorded in that FY. The "cumulative" numbers include the interactions from the program beginning in FY02 to that current year. The "active" numbers only include interactions by authors who are either active in that given FY or were active in a prior and later FY.

Table 3 C&N CTA: Number of vertices and edges

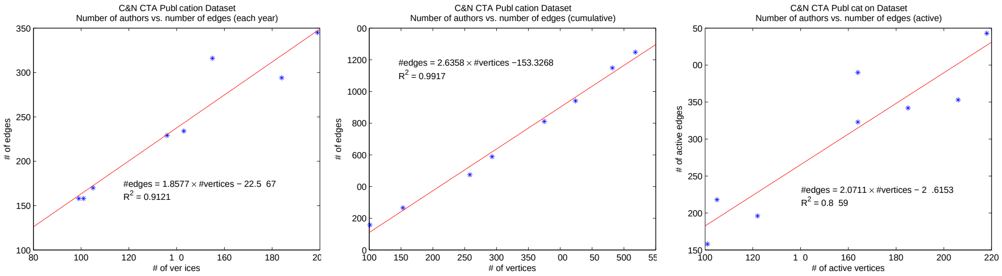
Fiscal Year	Per Year: # Authors	# Edges	Cumulative: # Authors	# Edges	Active: # Authors	# Edges
2002	101	158	101	158	101	158
2003	99	158	153	266	122	196
2004	184	294	258	475	206	353
2005	143	234	293	589	185	342
2006	199	345	375	811	218	443
2007	136	229	424	941	164	323
2008	155	316	482	1149	164	390
2009	105	170	518	1248	105	218

There is strong empirical evidence that a linear relationship exists between the number of vertices and the number of edges for the cumulative network. The evidence is weaker for the yearly and active networks. The best linear unbiased estimator for the cumulative network is given by

$$\#\text{edges} = 2.6358 \times \#\text{vertices} - 153.3268, \tag{3}$$

with a nearly unitary coefficient of determination of $R^2 = 0.9917$.

The slope and intercept estimates and the coefficients of determination for simple linear regressions on the dataset for each network are given in Fig. 3. The greater slope and lower intercept estimate for the cumulative data (Fig. 3b) compared with the per-year data (Fig. 3a) indicates that the proportion of new edges each year is greater than the proportion of new vertices each year. This perhaps implies that new authors were collaborating with different existing authors and/or existing authors were forming new collaborations among themselves. This is an expected observation given the program goals of fostering collaboration, which leads to more interactions and/or larger group interactions (see Fig. 1). As each new author joins the program, that author will publish with more than 1 author on average, thereby generating multiple new edges. These new authors are usually new students, post-docs, and external collaborators as opposed to new PIs in the program.



(a) Per Year (b) Cumulative (c) Active

Fig. 3 The C&N CTA: Number of vertices vs. number of edges

At the same time, the longer an author remains in the program, the probability of that author developing a new collaboration with another author already in the program increases. Note that the negative intercept in a linear relationship such as those in Fig. 3b indicates a generally increasing trend of authors who have a short lifetime in the program (Table 4), thereby prohibiting collaborations (edges) with other (later or continuing) program authors. This is a consequence of a program ending, which shifts the line down.

Table 4 C&N CTA: New authors vs. single-year authors

Fiscal Year	2002	2003	2004	2005	2006	2007	2008	2009
New authors	101	52	105	35	82	49	58	36
Single-year authors	31	12	45	23	42	27	47	36

Whereas a linear fit appears to fit well to the yearly and cumulative data, a linear model is a poorer match for the active vertices and edges in the dataset (Fig. 3c).

2.1.2 Components

The number of components of the C&N CTA network over the course of the program is detailed in Table 5. In any given year, there are time constraints that limit the number of collaborations that will result in a publication. Authors form core groups that may or may not interact with other groups within a given year or given sequence of multiple years. It at first appears surprising that the number of components did not markedly decline over the evolution of the cumulative network. However, this is explained by comparison with the active network components. Half the components of the cumulative network are composed of inactive authors (vertices).

Table 5 C&N CTA: Number of components

Fiscal Year	2002	2003	2004	2005	2006	2007	2008	2009
# of components (p.y.)	20	15	31	21	26	16	17	18
# of components (cum.)	20	21	27	19	20	18	16	16
# of components (act.)	20	21	28	16	15	8	8	7

The sizes of the key components and median size are presented in Table 6. The primary and secondary components (or largest and second largest components) of the yearly network contain, respectively, 34.8 and 16.9 vertices on average with small deviations from these averages. These components represent the collaborative activity in that given year, due to the various task collaborations. The active network contains more variability with larger secondary components, indicating that the connections between various components are only preserved in the cumulative network by inactive authors.

Table 6 C&N CTA: Sizes of components (measured by # of authors)

Fiscal Year	2002	2003	2004	2005	2006	2007	2008	2009
Primary component size (p.y.)	40	32	23	31	48	35	44	25
Secondary component size (p.y.)	5	10	17	17	28	21	19	18
% in primary component (p.y.)	39.6	32.3	12.5	21.7	24.1	25.7	28.4	23.8
Mean component size (p.y.)	5.1	6.6	5.9	6.8	7.7	8.5	9.1	5.8
Median component size (p.y.)	3	4	4	5	4	3.5	5	3.5
Primary component size (cum.)	40	57	110	175	290	347	419	453
Secondary component size (cum.)	5	15	18	35	10	9	9	9
% in primary component (cum.)	39.6	37.3	42.6	59.7	77.3	81.8	86.9	87.5
Median (cum.)	3	4	4	5	4	4	4	4
Primary component size (act.)	40	37	44	59	172	85	76	52
Secondary component size (act.)	5	12	29	48	10	63	39	34
% in primary component (act.)	39.6	30.3	21.4	31.9	78.9	51.8	46.3	49.5
Median (act.)	3	3	4	5	3	3.5	9.5	4

By allowing interactions of inactive authors to persist, the cumulative network is characterized by a dominating primary component with a relatively shrinking secondary component. This behavior has been previously characterized for much larger networks.[5] The only restriction on the growth of the primary component is the number of inactive components. These small inactive components become inactive due to the relatively short lifetimes of their authors' activity. Six of the 15 nonprimary components in the cumulative network consist entirely of authors who were only active in the network for a single FY, i.e., their papers are only listed on a single year's spreadsheet. For whatever reason, these 17 authors were active in the program too briefly for collaboration to emerge with authors in the primary component. The primary component also has a large number of authors who have short "lifespans" in the network: 49.89% of these vertices were only active in a single year compared with 53.85% in the small components. However, only 81.90% of authors in the primary component have lifespans of at most 3 years compared with 90.77% of authors in the small components. The slightly larger percentage of authors who have much shorter lifespans also limits the opportunities for a collaborative link to be established that would potentially reduce the number of components.

Unfortunately, the size of the C&N CTA dataset is insufficient to determine if the distribution of the component sizes has any pattern such as a power law, as has been found in other networks.[11]

2.1.3 Distances

2.1.3.1 Eccentricity

For the C&N CTA dataset, the center subgraph for the final year of the cumulative network consists of 2 disconnected vertices with eccentricity 7 corresponding to John S Baras and ND Sidiropoulos. The periphery vertices have eccentricity 12 and consist of 26 vertices that induce a 17-component graph with 2 triangles, 5 edges, and 10 isolated vertices. Eccentricity is one way to measure the "centrality" of a vertex with respect to the rest of the graph (component). Obviously, the more central a vertex is the more influential an author is, as the graph is in some sense built upon the foundation of the center authors' collaborations. This interpretation is especially valid here since these authors existed early in the evolution of the (cumulative) network. (Unlike John S Baras, ND Sidiropoulos does not participate in later years of the program and, hence, does not seem as important in the yearly or active networks.) Collaborations of authors on the periphery, e.g., interactions to leaf vertices, are remote from the core collaborations in the graph. Taking this analogy further, this means that the small non-primary components are even more remote.

2.1.3.2 Primary Component

The primary component for the cumulative network does exhibit the small world phenomenon, in that the typical path length is $\log(n)$, where n is the number of authors in the component. Figure 4 shows the average and median path lengths in the graph, as well as its diameter, corresponding to data in Table 7.

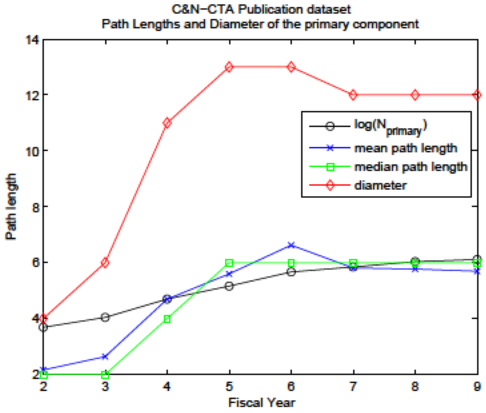


Fig. 4 The C&N CTA: Path lengths and diameter over the evolution of the cumulative network

Table 7 C&N CTA: Primary component path length characteristics

Fiscal Year	2002	2003	2004	2005	2006	2007	2008	2009
$\log(n_{\text{primary}})$	3.69	4.04	4.70	5.16	5.67	5.85	6.04	6.12
Mean path length	2.17	2.65	4.70	5.60	6.63	5.81	5.77	5.70
Median path length	2	2	4	6	6	6	6	6
Diameter	4	6	11	13	13	12	12	12

2.1.4 Clustering

The global clustering coefficient (transitivity) and average local clustering coefficient, both per year and cumulatively, are shown in Fig. 5.

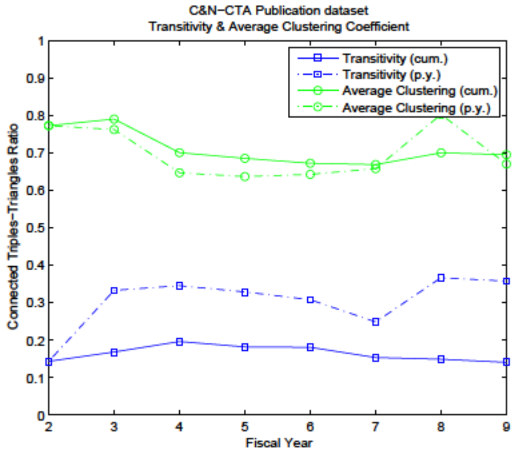


Fig. 5 The C&N CTA: Graph transitivity and average local clustering coefficient

Transitivity for the C&N CTA network graph for each FY (cumulatively) does not vary much from its mean, and the yearly graph only varies slightly more after the initial increase in the second year of the program. The sample mean and variance of the average local clustering coefficient (yearly and cumulatively), as well as the transitivity, are shown in Table 8.

Table 8 C&N CTA: Transitivity and average clustering coefficient

	Mean	Variance
Transitivity (cum.)	0.1645	0.0004
Transitivity (p.y.)	0.3043	0.0056
Average Clustering Coefficient (cum.)	0.7113	0.0021
Average Clustering Coefficient (p.y.)	0.6996	0.0046

For comparison, the average local clustering coefficient for an Erdös-Rényi graph approaches the ratio of the average degree to the graph size, which is also the probability of an edge existing between 2 vertices.[12] In the case of the C&N CTA co-authorship graph, the ratio of edges to possible edges is an order of magnitude less than the numbers in the table. This is evidence of the small-world nature of this co-authorship graph, which is perhaps not surprising given that the network is constrained to participation within the C&N CTA.

If authors who formed connected triples but not triangles are no longer in the program, then the opportunity to "close" the triple and form a triangle is lost. Considering the large number of authors with short lifetimes in the network, the large clustering coefficients in Table 8 indicate that clustering was an immediate and, therefore, essential characteristic of the collaborations.

2.2 Centrality

2.2.1 Vertex Degree

2.2.1.1 Local Characteristics of the Vertex Degrees

For the C&N CTA network, the most prominent authors in terms of various centrality metrics (vertex degree, closeness, and betweenness) are given in Table 9.

Giannakis, by the degree metric alone, is clearly a superhub relative to the other hubs on this list. However, as degree only determines the local connections of an author, this only implies that Giannakis collaborated with a large number of authors. This is not surprising as we have seen from Table 2 that he collaborated on a significant number of papers. It is also known that Giannakis had a large number of students that participated in the program.

The list of authors with high degree in Table 9 does not give a sense of how relatively large these degrees are to the other vertices. For comparison, the mean and median degrees for the yearly network, the active network, and the cumulative network (both the whole network and just its primary component in this latter case) are given in Table 10. It makes sense to compare with the mean and median degrees of the vertices in the primary component since all the authors listed in Table 9 are in the primary component. In fact, the first 64 authors ranked by vertex degree are in the primary component. Since the degrees have a heavy-tailed distribution, the mean is biased high. Even still, the program's high-degree authors in Table 9 have

Table 9 C&N CTA: Centrality top-20 lists

Rank	Degree		Closeness		Betweenness	
1	Georgios B Giannakis	68	ND Sidiropoulos	0.2684	John S Baras	0.3324
2	John S Baras	32	Ananthram Swami	0.2681	Georgios B Giannakis	0.3264
3	Lang Tong	29	Brian M Sadler	0.2599	Brian M Sadler	0.3140
4	Xiaoli Ma	26	Georgios B Giannakis	0.2592	ND Sidiropoulos	0.3007
5	Anthony J McAuley	26	Tao Jiang	0.2482	Tao Jiang	0.2987
6	Ananthram Swami	25	John S Baras	0.2383	Ananthram Swami	0.2162
7	Brian M Sadler	24	John E Kleider	0.2353	Xiaodong Cai	0.1250
8	Myung Jong Lee	23	Shengli Zhou	0.2348	Yi Sun	0.1221
9	Daniel Sterne	23	Lang Tong	0.2346	Myung Jong Lee	0.1218
10	Tarek N Saadawi	22	Xue Wu	0.2346	Lang Tong	0.1130
11	Sergio Verdu	22	Steve Gifford	0.2331	H Vincent Poor	0.1095
12	Mariusz A Fecko	20	Xiaoli Ma	0.2307	Yingbo Hua	0.1011
13	Yingbo Hua	20	Qing Zhao	0.2307	Giovanni Di Crescenzo	0.1010
14	Qing Zhao	20	Liuqing Yang	0.2269	Lili Huang	0.0945
15	Richard Gopaul	19	X Liu	0.2266	Radha Poovendran	0.0938
16	Kyriakos Manousakis	19	A Stamoulis	0.2253	Qing Zhao	0.0936
17	Radha Poovendran	19	J Tao	0.2252	Mariusz A Fecko	0.0915
18	Adarshpal S Sethi	19	Yingbo Hua	0.2249	Kyriakos Manousakis	0.0891
19	Errol L Lloyd	18	Zhengyuan Xu	0.2249	Youngchul Sung	0.0869
20	Latha Kant	17	M Ghogho	0.2243	Maria Striki	0.0841
20	Sunil Samtani					

at least 4 times the end-of-program mean degree (4.8185). These authors also have at least 7 times the end-of-program median degree (3).

Table 10 C&N CTA: Mean and median degrees

Fiscal Year	2002	2003	2004	2005	2006	2007	2008	2009
mean (p.y.)	3.1287	3.1919	3.1957	3.2727	3.4673	3.3676	4.0774	3.2381
mean (act.)	3.1287	3.2131	3.4272	3.6973	4.0642	3.9390	4.7561	4.1524
mean (cum.)	3.1287	3.4771	3.6822	4.0205	4.3253	4.4387	4.7676	4.8185
mean (cum. p.c.)	4.2000	4.0351	4.3636	4.4800	4.6621	4.8761	5.1456	5.1876
median (p.y.)	3.0000	2.0000	2.0000	3.0000	3.0000	2.0000	3.0000	3.0000
median (act.)	3.0000	2.0000	3.0000	3.0000	3.0000	3.0000	4.0000	3.0000
median (cum.)	3.0000	3.0000	3.0000	3.0000	3.0000	3.0000	3.0000	3.0000
median (cum. p.c.)	3.0000	3.0000	3.0000	3.0000	3.0000	3.0000	4.0000	4.0000

Table 10 also reflects that the mean degree over the evolution of the network does not appear to have stabilized by the end of the program. This appears to be due to the increasing trend of the mean degree in each FY with a rather sizable jump from FY07 to FY08. The median degree appears more stable. Even the jump in the median degree from FY07 to FY08 in the yearly or active network is not sufficient to change the median degree in the cumulative network, although it does imply some incentive for the authors to become more collaborative late in the program.

2.2.1.2 Global Characteristics of the Vertex Degrees

The distribution for the C&N CTA final year cumulative network is shown in Fig. 6. It is well documented[5,10,13] that the degree distribution of co-authorship graphs tend to roughly follow a power law, possibly with an exponential cut-off in the tail and a hook at the head. It is surmised that the hook is a characteristic of measuring the new authors who join the network over a finite timeline, and the power law effect is evidence of well-established authors who find it is easier to establish collaborations the more they publish. Since many studies have occurred over a short timeframe, it is doubtful that the data captures the full careers of most of the authors, which leads to some question over the true nature of the long-term evolutionary degree distribution. However, analyses of short-term datasets have been consistent in characterizing the degrees of scientific co-authorship, as in the case here.

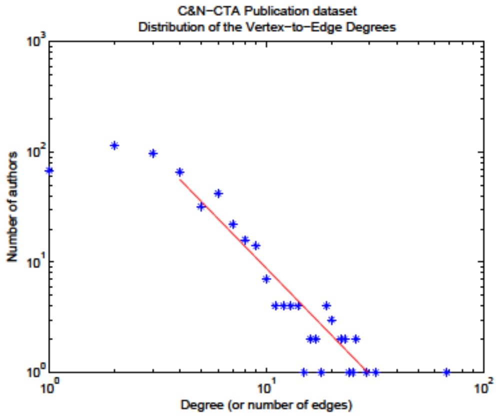


Fig. 6 The C&N CTA co-authorship network: Vertex degree distribution (exponent = -2.02**)**

The distribution shows a clear hook at the head of the distribution and some evidence of a power law in the tail. Because of the small size of this network, the tail is noisy and it is more difficult to infer if the distribution has a true power law tail. Indeed, if the network were larger and the noise tempered, then this distribution would likely be similar to that observed in the literature. This is somewhat surprising since this network is constrained to a particular program, hence omitting collaborations external to the program. Ignoring extremal values, the distribution is estimated to satisfy

$$\text{\#authors} = 924.91 \times \text{degree}^{-2.02}. \tag{4}$$

2.2.2 Closeness

The authors with large closeness indices (for the primary component of the end-state cumulative network) are listed in Table 9. Perhaps the most surprising result is that the highest ranked author in terms of closeness centrality, ND Sidiropoulos, is not ranked high in terms of degree centrality or even in the number of publications. Sidiropoulos has a relatively low degree of 9 (ranked 57th) and 14 publications (ranked 36th). Of the top-20 ranked authors, closeness centrality shares 8 authors with the degree centrality rank list and 9 authors with the publication rank list. However, Sidiropoulos is 1 of the 2 vertices in the center (along with John S Baras who is ranked 6th in closeness).

2.2.3 Betweenness

Betweenness for each author in the primary component of the end-state cumulative network is listed in Table 9. Only 7 authors appear on all 3 top-20 lists in Table 9: John S Baras, Georgios B Giannakis, Brian M Sadler, Ananthram Swami, Lang Tong, Yingbo Hua, and Qing Zhao. Each of these also appears in the top-20 list for publication rank in Table 2.

The Pearson and Spearman correlation coefficients comparing the primary component authors' publication and centrality measures are given in Table 11.

Table 11 C&N CTA: Pearson and Spearman correlation values for several centrality measures

		Publications	Degree	Closeness	Betweenness
Pearson	Publications	*	0.8268	0.3303	0.6147
	Degree	0.8268	*	0.3086	0.6766
	Closeness	0.3303	0.3086	*	0.4028
	Betweenness	0.6147	0.6766	0.4028	*
Spearman	Publications	*	0.5038	0.2469	0.7503
	Degree	0.5038	*	0.1925	0.7119
	Closeness	0.2469	0.1925	*	0.3111
	Betweenness	0.7503	0.7119	0.3111	*

By Pearson correlation, an author's number of publications is most correlated with the number of collaborators, whereas closeness centrality and the number of collaborators exhibit the least correlation. By Spearman's rank correlation, the highest correlation occurs between the number of publications and the betweenness centrality of an author. Again, the least correlation occurs between closeness and vertex degree.

3. Simplicial Complex Analysis of the C&N CTA Publication Network

3.1 Motivation

Graphs are useful models of pairwise relations between actors but require significant modification to represent group relations efficiently. For example, a clique between a vertices in a graph, represented as $\binom{a}{2}$ edges, does not distinguish between $\binom{a}{2}$ independent pairwise relationships versus 1 collective relationship among the vertices. In the co-authorship context, a clique of 3 authors in a graph could mean that the authors wrote (at least) 1 paper collectively or that each author wrote (at least) 1 paper with each of the other 2 separately.

There are several approaches to address the above issue. The first option is to use a bipartite graph (or affiliation network). This is the most complete representation since it encompasses all the relationships, with multiplicities, that exist in the network. These networks are often considered difficult for analysis, and so "one-mode projections" are often used instead.[14] The typical co-author network is derived from such a projection. More generally, a hypergraph can be a better representative projection as it captures the group structures. Because of the topological and combinatorial properties of simple hypergraphs with the subset closure property, we represent the C&N CTA dataset as a simplicial complex.

3.2 Definition and Background

A more complete introductory description of simplicial complexes can be found elsewhere,[15,16] only some of which is detailed here.

A *hypergraph* $\mathcal{H} = (\mathcal{V}, \mathcal{E})$ is simply an extension of a graph that allows for any multiple of vertices to exist in an "edge," now called a *hyperedge*. Formally, the vertex set $\mathcal{V}$ is exactly as described in the definition of a graph, but now the edge set consists of sets $\epsilon \in \mathcal{E}$ of any size such that $\epsilon \subset \mathcal{V}$. This allows for characterizations beyond pairwise relations found in a simple graph. An *abstract simplicial complex* Δ is a special case in which every subset of a hyperedge is also a hyperedge. Because of this closure property on subsets, simplicial complexes are amenable to mathematical formalism in combinatorics, abstract algebra, and topology (see Munkres[16] for definitions and properties relating to simplicial complexes and homology not presented here).

More formally, an abstract simplicial complex is a collection Δ of sets such that every subset of a set in the collection is also in the collection, i.e., if $\sigma \in \Delta$ and $\tau \subset \sigma$ then $\tau \in \Delta$. An element of Δ is called a *simplex*. The union of all the simplices (sets) in Δ is the vertex set $\mathcal{V}$, and the elements of the union are the vertices of the complex. Note that this approach of defining a simplicial complex first describes the larger structure and then describes its components, although we could have started with a description of the smallest structures (the vertices) and built up to the largest (the complex) as is often done in the description of a graph.

By definition, every simplex is a set of vertices and, hence, can be denoted by the vertices it contains, e.g., $\sigma = v_0 v_1 \ldots v_k$. The *dimension* of the simplex σ is 1 less than the number of its vertices, i.e., $\dim(\sigma) = |\sigma| - 1$. A simplex of dimension k is called a k-*simplex*. A proper subset of the simplex σ is called a *face*. Every face is also a simplex. A face of dimension m is called an m-*face*. If a simplex is not a face of any other simplex in the complex, then it is called a *facet* (or *maximal simplex*).

The dimension of the complex is the supremum of the dimensions of its simplices.

For simplicity, we assume a finite vertex set, so that $\dim(\Delta) < \infty$. A *subcomplex* of Δ is an abstract simplicial complex Θ such that every simplex in Θ is also in Δ. The k-*skeleton* of Δ is the subcomplex that includes every simplex of Δ of dimension at most k. In particular, the 1-skeleton is often referred to as the underlying graph of the complex, though technically the 1-skeleton is isomorphic to a graph. The simplicial complex can be described by a list of all of its simplices, called a *vertex scheme*, or simply by a list of its facets since we can assume the subsets of the facets are implicitly in the list.

For any abstract simplicial complex, there exists a geometric realization (simply called a *simplicial complex*) satisfying certain properties. Among these properties, a k-simplex is the convex hull of $k+1$ affinely independent points and is, thus, k-dimensional. For example, if $v_0, v_1, \ldots, v_k$ are affinely independent, then the simplex of these points is simply the set

$$\left\{ p : p = \sum_{i=0}^{k} \alpha_i v_i, \sum_{i=0}^{k} \alpha_i = 1 \right\}. \tag{5}$$

Each of these points defining the simplex is a vertex. Hence, a 0-simplex is a vertex (point), a 1-simplex is an edge (line segment between 2 vertices), a 2-simplex is a triangle, a 3-simplex is a tetrahedron, a 4-simplex is a pentachoron, and so on. A face of a simplex is the convex hull of a proper subset of the vertices in the simplex. Then a simplicial complex is a “gluing” together of simplices such that the following is true:

1. Any face of a simplex in the simplicial complex is also in the complex and

2. The intersection of any 2 simplices is a face of both simplices.

This latter condition requires that the geometric realization reside in a dimension of sufficient size (see Section 3.5.4). A geometric realization can be useful for visualizing the relationships between the vertices.

3.2.1 Homology

An *oriented simplex* is a simplex with an orientation, denoted $[v_0, v_1, \ldots, v_k]$, such that 2 oriented simplices with the same vertices are equivalent if they differ by an even permutation of their orderings. This induces a sequence of modules (or free Abelian groups) over a ring, denoted $\mathcal{C}_k(\Delta)$, with the basis set of (oriented)

k-simplices of Δ for each k. The elements of $\mathcal{C}_k(\Delta)$ are called *k-chains* and are written as $\sum_i a_i \sigma_i^{(k)}$, where each a_i is an element in the ring and each $\sigma_i^{(k)}$ is an oriented simplex.[17] (Every k-simplex in Δ is a generator in the group. Also, by definition, $\mathcal{C}_k(\Delta) = 0$, the trivial group, for every $k < 0$ and $k >$ dim(Δ).) If $\sigma_j^{(k)}$ is a simplex with the same vertices as those in $\sigma_i^{(k)}$ but with an ordering differing by an odd permutation, then $\sigma_j^{(k)} = -\sigma_i^{(k)}$.

The sequence of modules over Δ is connected as a *chain complex* by *boundary operators* (homomorphisms) $\partial_k : \mathcal{C}_k(\Delta) \to \mathcal{C}_{k-1}(\Delta)$ defined by

$$\partial_k[v_0, v_1, \ldots, v_k] = \sum_{j=0}^{k} (-1)^j [v_0, v_1, \ldots, \hat{v}_j, \ldots, v_k], \tag{6}$$

where $\hat{v}_j$ denotes that this vertex is missing from the oriented simplex. That is, the kth boundary operator maps every k-simplex to a sum of its oriented $(k-1)$-faces, i.e., a sum of the simplices in its *boundary*. (We define ∂_k to be the zero map if $k < 1$ or $k >$ dim(Δ).) Hence, we refer to the image of ∂_{k+1} as the *subgroup of k-boundaries in $\mathcal{C}_k(\Delta)$* and denote it as $\mathcal{B}_k(\Delta)$. If for a k-chain $c^{(k)} \in \mathcal{C}_k(\Delta)$ we have $\partial_k c^{(k)} = 0$, then we say that $c^{(k)}$ is a k-cycle (analogous to cycles in a graph, where the “flow” into each vertex equals the “flow” out of that vertex). We refer to the kernel of ∂_k as the *subgroup of k-cycles in $\mathcal{C}_k(\Delta)$* and denote it as $\mathcal{Z}_k(\Delta)$. It is trivial to show that $\mathcal{B}_k(\Delta) \subset \mathcal{Z}_k(\Delta)$ for every k. Therefore, we have the *kth homology group* of Δ defined as the factor group

$$\mathcal{H}_k(\Delta) = \mathcal{Z}_k(\Delta)/\mathcal{B}_k(\Delta). \tag{7}$$

A topological space can be classified (up to a homeomorphism) by the determination of its homology class, a topological invariant. For simplicial complexes, the homology groups for every k determine the classification. Of particular interest are the *Betti numbers*, the ranks of the homology groups of Δ, which generally characterize the number of unconnected k-dimensional surfaces. Formally, the kth Betti number is denoted by

$$b_k(\Delta) = \text{rank}(\mathcal{H}_k(\Delta)). \tag{8}$$

Specifically, $b_0(\Delta)$ is the number of connected components of Δ, $b_1(\Delta)$ is the number of 2-dimensional “holes” in Δ, $b_2(\Delta)$ is the number of 3-dimensional holes or voids in Δ, etc.

3.2.2 Laplacian Matrices

The *kth combinatorial Laplacian operator* is the endomorphism $\mathcal{L}_k$ on $\mathcal{C}_k(\Delta)$ defined by $\mathcal{L}_k = \partial_k^* \circ \partial_k + \partial_{k+1} \circ \partial_{k+1}^*$ where ∂_k^* is the adjoint of ∂_k. It can be shown[18] that if the ring in $\mathcal{C}_k(\Delta)$ is $\mathbb{R}$, for example, then

$$\ker(\mathcal{L}_k(\Delta)) \cong \mathcal{H}_k(\Delta). \tag{9}$$

For each k, given an ordering of the n_k k-simplices of a complex, the Laplacian operator has an equivalent representation on $\mathbb{R}^{n_k}$:

$$\boldsymbol{L}^{(k)} = \boldsymbol{B}^{(k)T}\boldsymbol{B}^{(k)} + \boldsymbol{B}^{(k+1)}\boldsymbol{B}^{(k+1)T}, \tag{10}$$

where $\boldsymbol{B}^{(k)} : \mathbb{R}^{n_k} \to \mathbb{R}^{n_{k-1}}$ is the *kth boundary matrix* defined by the boundary operator in Eq. (6) and the ordering of the k- and $(k-1)$-simplices. Note that $\boldsymbol{L}^{(0)}$ is the familiar graph Laplacian, where $\left|\boldsymbol{B}^{(1)}\right|$ is the vertex-edge incidence matrix. Also, the *kth combinatorial Laplacian matrix* is the sum of positive definite matrices, and if $\boldsymbol{x} \in \text{null}(\boldsymbol{L}^{(k)})$, we have that $\boldsymbol{x}$ corresponds to a k-cycle that is orthogonal to the boundary space. It is not true, however, that each such k-cycle will exist in null$(\boldsymbol{L}^{(k)})$.

3.2.3 Representing Collaborations as a Simplicial Complex: The C&N CTA Dataset

For the C&N CTA dataset described in Section 1, or for any co-authorship network dataset, a simplicial complex can be generated by identifying vertices in the complex with authors and identifying simplices with the co-authorship relation, i.e., if the authors corresponding to a set of vertices collaborated on (co-authored) a publication in the dataset, then the simplex of those vertices exists in the complex. This definition using collaboration ensures the existence of a simplicial complex since the subset closure property holds. A visualization of the C&N CTA simplicial complex is shown in Fig. 7.

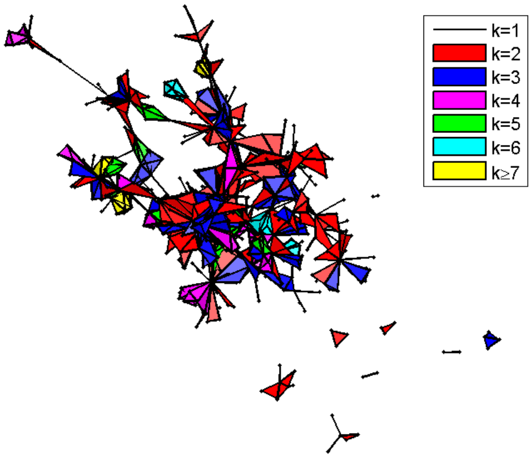


Fig. 7 The C&N CTA simplicial complex

3.3 Connectivity

The key results and conclusions from this section include the following:

1. The subset property in complexes means the number of simplices is not an accurate model of the number of collaboration groups.
2. The number of facets represents the number of distinct publishing groups (often, however, there exists a core publishing subgroup of the group).
3. The sample size may be too small to determine a model for the yearly network links; the cumulative network model is smooth enough to permit the construction of a model.
4. There is a strong correlation between the number of k-simplices and number of $(k+1)$-simplices (and little to none between k-simplices and facets) in the cumulative model.
5. There is a strong linear relationship between the number of facets and the number of papers.

3.3.1 Network Size

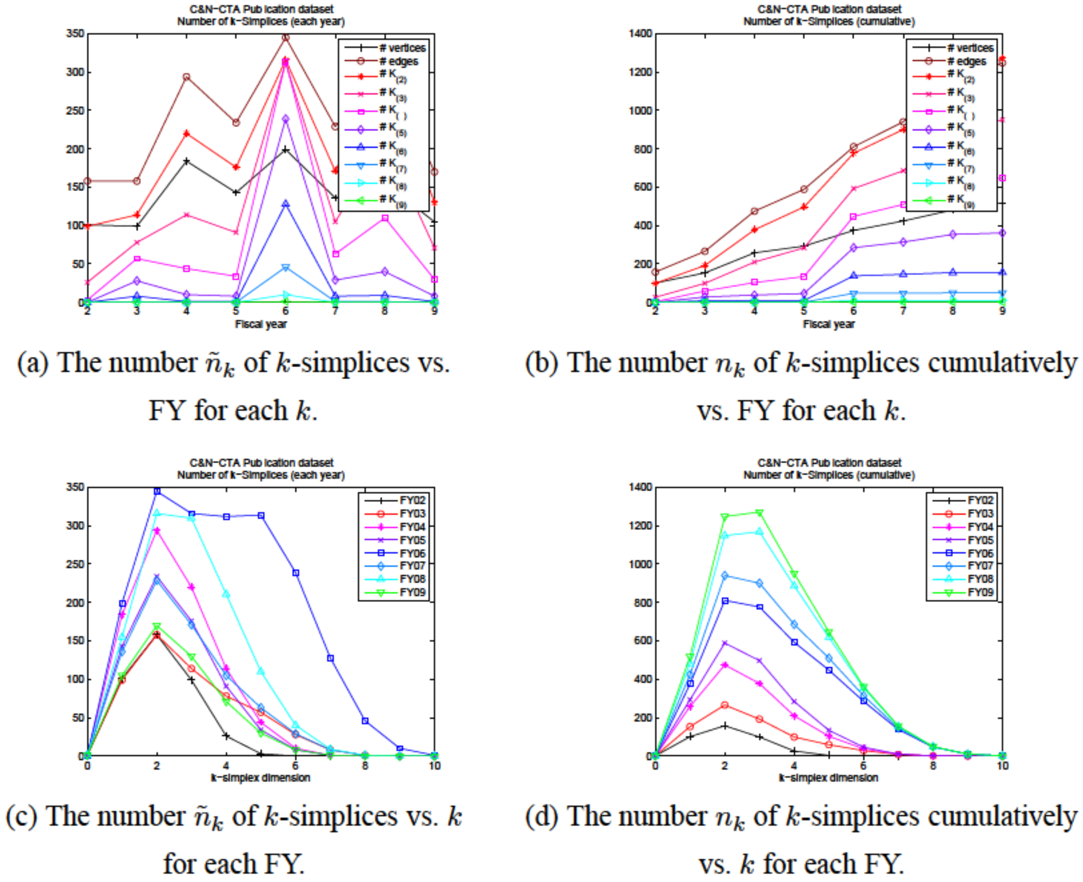
The evolution (or growth) of the number of vertices and edges is covered in Section 2.1.1. With a network simplicial complex, there are higher-dimensional objects that characterize the network's evolution. The obvious objects to measure are the k-dimensional simplices and facets for $k > 2$, i.e., the co-authorship publishing subgroups and groups in the dataset.

Given the facet or simplex list (for each year), it is a simple counting problem to determine the number of objects in each dimension. If the boundary matrices have been constructed then the number of k-simplices corresponds to the number of columns of $\boldsymbol{B}^{(k)}$ or the number of rows of $\boldsymbol{B}^{(k+1)}$, and the number of facets can also be found from the boundary matrices by determining the simplices that are not part of the boundary of a higher-dimensional simplex.

3.3.1.1 k-Simplices

The numbers of k-simplices, for different k, for the yearly and cumulative networks[19] are shown in Fig. 8 against the FY and the dimension k. We use $\tilde{n}_k$ and n_k to denote the number of k-simplices yearly and cumulatively, respectively. The plot in Fig. 8a demonstrates significant variability each year for each dimension. This may be due to correlation with the number of papers produced, which we discuss later. A clear inference from the figure is that the $\tilde{n}_k$ appear to be rank correlated, i.e., if $\tilde{n}_{k,y_i} > \tilde{n}_{k,y_j}$ (where the y_i index denotes the ith FY), then generally $\tilde{n}_{l,y_i} > \tilde{n}_{l,y_j}$, as well, regardless of l and k. (The cases where $l = 0$ and $l = k + 1$ are discussed in the next section.) This explains the similar directional shifts in the lines in Fig. 8a are similar.

Figure 8b demonstrates the smoothing effect when considering the cumulative network, as the effect of new or active collaborations is tempered by existing collaborations from prior years. Each line in the plot is monotone increasing because each year brings new collaborations and no collaborations are removed. Also note that the n_k are generally sortable (i.e., there are more k-simplices than $(k + 1)$-simplices), with the exceptions of the vertices and the change-point occurring in FY06. The primary reason for this is the closure property of subsets in simplicial complexes determines the rate of growth in the number of k-simplices when simplices are created among a single existing vertex and new vertices. However, when simplices are created among existing vertices, which may already have connections,



(a) The number $\tilde{n}_k$ of k-simplices vs. FY for each k.

(b) The number n_k of k-simplices cumulatively vs. FY for each k.

(c) The number $\tilde{n}_k$ of k-simplices vs. k for each FY.

(d) The number n_k of k-simplices cumulatively vs. k for each FY.

Fig. 8 The number of k-simplices vs. FY and the number of k-simplices vs. simplex dimension, for the yearly and cumulative networks

this should affect the rate of growth. For example, we can see that n_3 (the number of tetrahedrons or groups of 4) is on track to pass n_0 at FY05, and this indicates that 3-simplices are being created at a faster rate than vertices, which must be partly because of new simplices (collaborations) among existing vertices (authors).

The change-point at FY06 in Fig. 8b demonstrates a peculiarity with the closure property of subsets in simplicial complexes. There is a significant jump in n_4 (5-author collaborations); however, this is almost entirely due to a single conference paper (paper ID 652) that has 10 co-authors. A paper with 10 co-authors generates in the complex a 9-simplex that contains $\binom{10}{5} = 252$ 4-simplices or distinct 5-author subgroups. This paper accounts for the spikes in many of the values of n_k for each dimension k in FY06 in Fig. 8a, accounts for the super-linear growth in Fig. 8b, accounts for the significantly different distribution of the number of k-simplices in Fig. 8c, and also alters the shape of the distributions for the cumulative evolution in Fig. 8d. The smaller spikes in FY04 and FY08 are similarly due to a larger number of papers with 4 co-authors (3-simplices) and 6 co-authors (5-simplices), respectively. Absent this single large collaborative paper, the growth in n_k appears linear with respect to the change in FY for each k.

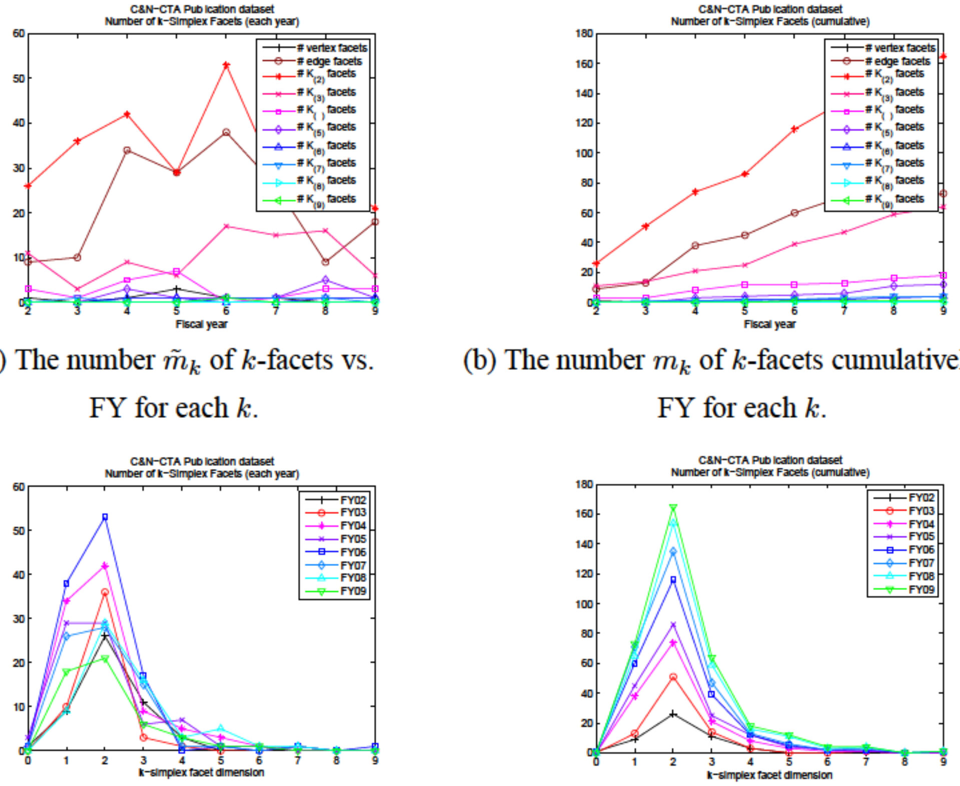
Figure 8c shows that the distributions of the sizes of the simplices are generally single-spiked with a mode at $k = 2$. That this is the mode is unsurprising given the distribution of the number of authors on a paper in Fig. 1. In most years, the papers consist primarily of 3-person collaborations (2-simplices). The 2 years where this distribution is significantly different are FY06 and FY08. In each case, the mode is still at $k = 2$, but there are papers with a larger number of collaborators, which sufficiently skew the distribution in the cumulative network in Fig. 8d. By the end of the program, the mode is shifted from $k = 2$ to $k = 3$.

One might argue that it is not desirable for a few papers (i.e., those few with the large collaborations) to have such a dramatic impact on the metrics for the growth of the complex or the distribution of its simplices. Hence, as a measure of the number of groups in a network, the number of simplices of a certain size should be viewed as a very loose upper bound. We consider the number of k-facets as an alternative metric in the next section.

3.3.1.2 k-Facets

The number of k-facets for each year and cumulatively are shown in Fig. 9 compared against the FY and the facet dimension k. We use $\tilde{m}_k$ and m_k to denote the number of k-facets yearly and cumulatively, respectively. As with the k-simplices, there is significant variability in the number m_k of facets for some dimensions in Fig. 9a. This variability is due to a large number of unique collaborations, primarily of dimensions $k = 1, 2, 3$. There are only a few large collaborations in this dataset and, therefore, they are a less interesting aspect of this figure. Another feature is that there appears to be less correlation between the facets of different sizes than evidenced in Fig. 8 (more on this in the next section). This is reasonable since there is no subset relation between facets of one size versus facets of another size.

Figure 9b demonstrates the smoothing of the variation seen in Fig. 9a over time. Note that the dramatic shifts seen in Fig. 8b at FY06 are not present here, since a single new unique collaborative effort (or paper) is only counted once regardless of its size. The general tendency is that the lines are increasing. But unlike the plots in



(a) The number $\tilde{m}_k$ of k-facets vs. FY for each k.

(b) The number m_k of k-facets cumulatively vs. FY for each k.

(c) The number $\tilde{m}_k$ of k-facets vs. k for each FY.

(d) The number m_k of k-facets cumulatively vs. k for each FY.

Fig. 9 The CN CTA co-authorship network: The number of k-simplex facets vs. FY and the number of k-simplex facets vs. facet dimension, for the yearly and cumulative networks

Fig. 8b, there are three instances where the number m_k decreases from one year to the next: twice for m_0 (FY03 and FY07) and once for m_1 (FY08). This occurs when a facet or group of authors (or a lone author) that already produced a publication produces a new work (collectively) with at least 1 other author.

Figure 9c plots the distributions of $\tilde{m}_k$ for each FY. Surprisingly, the distributions have a similar shape, i.e., a skewed distribution with a mode at $k = 2$. The variability is likely due to the extremely small sample size (there are significantly fewer facets than simplices in many simplicial complex representations). This modal feature is enhanced in the distribution shapes of m_k for each FY in Fig. 9d.

Facets are clearly a better representation of the number of distinct publishing groups. However, they mask the subgroup collaborations within each group. For example, a group of 3 co-authors may have jointly written several papers together and only

1 paper with another co-author. The facet count only includes the group of 4 co-authors and ignores that a face of that simplex was a significant contributor to the collaborative output. In light of this, we should view the number of k-facets as a lower bound of the actual publishing collaborations. To truly capture these groups and their output, weights on simplices are likely necessary, e.g., the weight of a simplex of authors could be the number of papers jointly published by those authors.[20]

As an aside, we also note that it is the intersection of these facets (publishing groups) that creates the connectivity necessary for this dataset to be represented as a simplicial complex instead of just a collection of simplices. This can be characterized by a generalization of the degree of a vertex, which is discussed in Section 3.3.3.2.

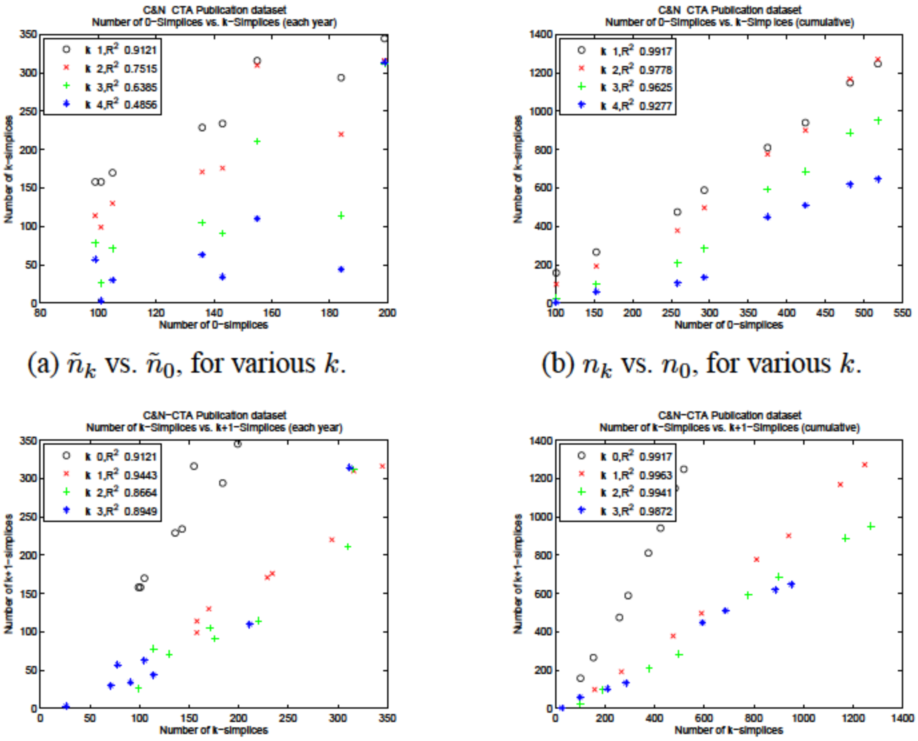
3.3.2 Correlation Among the Numbers of Simplices and Facets

An important characteristic studied in network graphs is the relationship between the number of vertices and the number of edges. This study extends in network simplicial complexes in 2 ways: 1) the relationship between the number of vertices and the number of k-simplices or k-facets, and 2) the relationship between the number of simplices/facets with dimension k and those with dimension $k + 1$.

3.3.2.1 k-Simplices

In Fig. 10a, we plot the number of k-simplices versus the number of vertices in each year's network and the coefficient of determination R^2 (or square of the Pearson correlation coefficient) for $k = 1, 2, 3, 4$. There is little evidence of any strong linear correlation each year between the number $\tilde{n}_0$ of vertices and the number $\tilde{n}_k$ of collaborations of a certain size. However, Fig. 10b does demonstrate an initially strong relationship between n_0 and n_k that gradually weakens as k increases. That the linearity develops in the cumulative observation of the network more so than the yearly observation is not surprising because of the embedded memory of including simplices from prior years. This might also indicate that the number of simplices for each year in the program may be too small a sample size for inferring a statistical relationship.

This hypothesis is provided more evidence in Fig. 10c, where there is weak evidence of a linear relationship between the number of k-simplices and the number of $(k + 1)$-simplices in the network each year. This is surprising because the closure property of simplicial sets guarantees $k + 1$ k-simplices in each $(k + 1)$-simplex.



(a) $\tilde{n}_k$ vs. $\tilde{n}_0$, for various k.

(b) n_k vs. n_0, for various k.

(c) $\tilde{n}_{k+1}$ vs. $\tilde{n}_k$, for various k.

(d) n_{k+1} vs. n_k, for various k.

Fig. 10 The number of k-simplices vs. the number of vertices and the number of $(k+1)$-simplices vs. the number of k-simplices, for the yearly and cumulative networks

One would assume that this property should dominate the statistic given a sufficient number of simplices; however, the connectivity between simplices sharing vertices and even k-faces can be sufficiently different (nonlinearly) from year to year, which affects the relationship between the number of k-simplices and $(k+1)$-simplices. This divergence from linearity may be a good sign if it indicates that groups intersect (share members). In the cumulative network statistics, there is a significantly greater number of simplices, and, with the embedded memory of prior years' simplices included, we do have clear evidence of a linear relationship in n_{k+1} vs. n_k, as shown in Fig. 10d. The sample linear regression models corresponding to this figure[21] are given as

$$n_{1,y} = 2.6358 n_{0,y} - 153.3268 \tag{11}$$

$$n_{2,y} = 1.0953 n_{1,y} - 111.1556 \tag{12}$$

$$n_{3,y} = 0.8177 n_{2,y} - 73.0945 \tag{13}$$

$$n_{4,y} = 0.7383 n_{3,y} - 29.3594 \tag{14}$$

It should be noted that the lowest R^2 when comparing $n_{k,y}$ against $n_{k-1,y}$ for $k = 1, \ldots, 9$ occurs when $k = 4$. However, the nonzero data points are scarce for higher dimensional simplices (larger k) in the yearly network and have infrequent changes in the cumulative network.

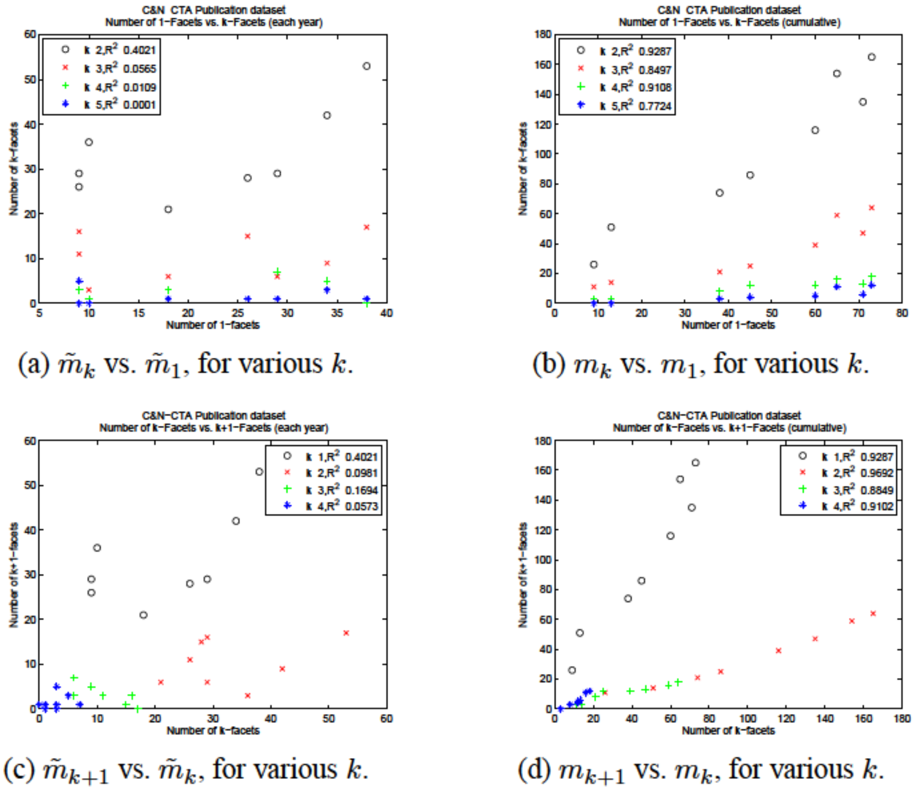
Regardless of the linear relationships between the simplices due to the activity within any given year, the (cumulative) network grows at a linear rate between the simplices. This linear relation in the growth of the network has important implications for the development of matching generative models.

3.3.2.2 k-Facets

Figures 11a and b compare the number of k-facets with the number of edge-facets, each year and cumulatively. Each year, the dataset is of insufficient size to infer a linear relationship between the number of edge-facets. Although the points might fit a linear model for the $k = 4, 5$ cases, they are well fit by a constant, which does not indicate any strong linear relation on $\tilde{m}_1$. (This is the implication of data that appears linearly flat with R^2 near zero.) The best linear fit in these 2 plots is for the cumulative network model for $k = 2$, but this is still a relatively weak fit based solely on R^2. However, since we lack any subset relationship when considering only the facets, the R^2 might be considered surprisingly high. Ultimately, a larger dataset is needed in order to permit definitive conclusions.

Figures 11c and d compare the number of $(k+1)$-facets with the number of k-facets, each year and cumulatively. Again, there is little evidence for any linear relationship for the network each year, and there is weak evidence of a linear relation for the cumulative network.

There is a higher correlation between the numbers of incident simplices than the number of incident facets, and there is a higher correlation for the cumulative network than the yearly network. That the number of incident simplices are more correlated is not a particularly surprising observation as every k-simplex must consist of $k + 1$ $(k - 1)$-simplices. The higher correlation for the cumulative network indicates that the new simplices entering the network might follow a pattern in how they connect to the existing simplices.



(a) $\tilde{m}_k$ vs. $\tilde{m}_1$, for various k.

(b) m_k vs. m_1, for various k.

(c) $\tilde{m}_{k+1}$ vs. $\tilde{m}_k$, for various k.

(d) m_{k+1} vs. m_k, for various k.

Fig. 11 The number of k-facets vs. the number of edge-facets and the number of $(k+1)$-facets vs. the number of k-facets, for each year of the network and cumulatively

3.3.3 Correlation Between the Number of Papers and Number of Simplices/Facets

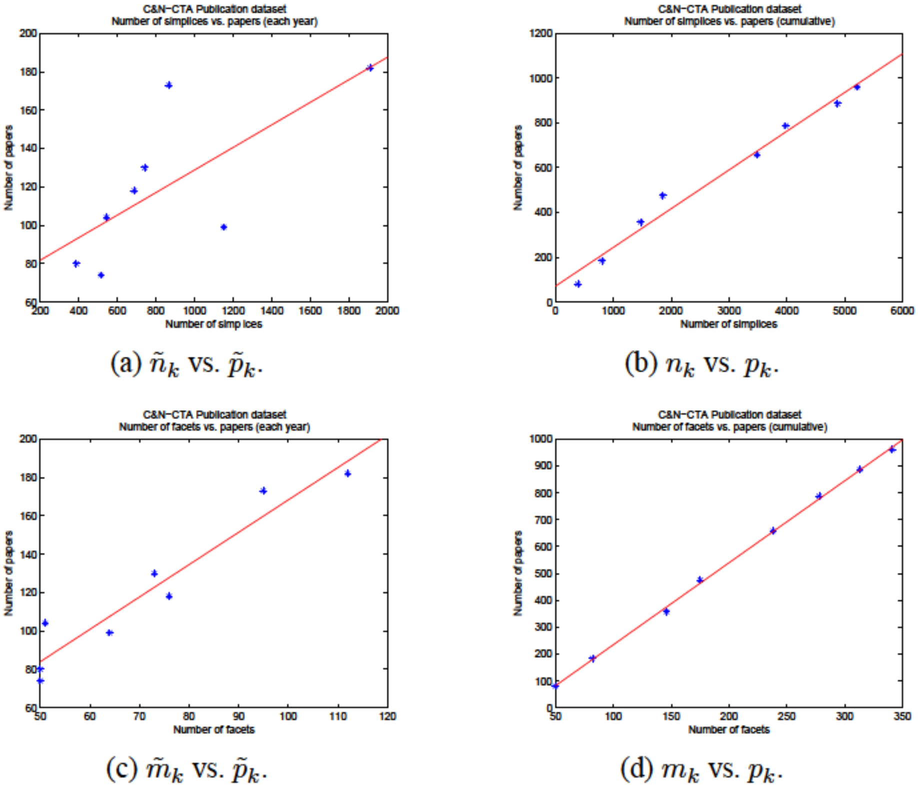
In graphs, there is often a linear relationship between the number of papers and the number of authors over time in a co-authorship network graph. This implies a number of properties of the co-authorship network, such as that the number of authors entering and leaving the network are constant and that there is an average rate of production for authors over a short time interval. In this section, we investigate if a linear relationship can be observed with respect to the number of groups of authors, thereby inferring analogous properties for groups.

3.3.3.1 k-Simplices

Figures 12a and b compare the total number of simplices (each year and cumulatively) with the number of papers. In the network data for each year, the coefficient $R^2 = 0.5174$, which indicates that the number of papers written in a given FY is not linearly dependent on the number of simplices that year. However, in the cumulative network, $R^2 = 0.9813$, which is a strong indication of a linear relationship, given by

$$p_y = 0.1731 \sum_k n_{k,y} + 70.6770. \tag{15}$$

This strong indication of a linear dependence is actually somewhat surprising. Different simplices are not necessarily distinct publishing groups. Just because a group of 3 authors collaborated on a paper does not mean that each pair collaborated on a separate paper. The same is true of a group of 4 authors. Yet a group of 3 authors contributes 7 simplices while a group of 4 authors contributes 15 simplices (ignoring potential intersections). Hence, a linear relationship between papers and simplices would only exist if on average groups of 4 authors are some unknown factor times more productive in the papers they produce than groups of 3 authors. Moreover, this unknown productivity factor would have to exist between groups of 5 authors (contributing 31 simplices) and groups of 4 authors as well.



(a) $\tilde{n}_k$ vs. $\tilde{p}_k$.

(b) n_k vs. p_k.

(c) $\tilde{m}_k$ vs. $\tilde{p}_k$.

(d) m_k vs. p_k.

Fig. 12 The number of k-simplices vs. the number of papers for each year of the network and cumulatively. The number of k-facets vs. the number of papers for each year and cumulatively

3.3.3.2 k-Facets

Figures 12c and d compare the total number of facets (each year and cumulatively) with the number of papers. The data for the year-by-year network exhibit a coefficient of determination of $R^2 = 0.9204$. The data for the cumulative network provide compelling evidence of a linear dependence since $R^2 = 0.9991$. The linear model is fitted by

$$p_y = 3.0515 \sum_k m_{k,y} - 70.8247. \tag{16}$$

Unlike the total number of simplices, which attributes greater weight to larger dimensional simplices because of the subset inclusion property, the total number of facets essentially characterizes the distinct working/publishing groups in the network. Hence, under the assumption of an average productivity among each group, facets are a more reliable indicator for the number of papers in the network.

3.4 Centrality

The key results and conclusions from this section include the following:

1. There is a clear power-law distribution for facet degree of a vertex. It is unclear if this extends to other simplex degrees.
2. The ranking of simplex degrees and other graph-based group centrality metrics do not appear correlated (using Spearman's coefficient).
3. The location of holes, i.e., the shortest cycles that do not bound, are incident to vertices with high graph centrality properties and form a component subcomplex.

For this section, we are only considering the complete network, i.e., the end state of the cumulative network.

3.4.1 Simplex Degrees

The degrees of the vertices were discussed earlier as 1 possible determination of vertex centrality in the graph by degree ranking as well as a characterization of a network by the distribution of the degrees. This graph degree distribution is still present in the simplicial complex (since the underlying graph is present). As such, the degree distribution is in many senses a model of the 1-skeleton. Similar models or characterizations might be possible (and, indeed, are necessary) for the higher-dimensional skeletons of the simplicial complex, which contain more dimensionality than just vertices and edges. In fact, this idea has been suggested before,[22] although the particular model analyzed there did not provide significant insight beyond the degree distribution.

The notion of degree of a vertex in a graph is generalizable to k-simplices. For example, the number of triangles incident to an edge is important in determining how many other actors a pair is jointly related to or, in the collaboration context here, how many other authors that a pair of co-authors jointly published a paper. This information is lost in a simple graph model of the collaborations and is an alternative to the number of co-authors with whom either author in a pair published a paper, a graph-based notion of group degree defined in Eq. (19). This induces a centrality characteristic based on ranking the productivity of pairs of authors in the C&N CTA dataset. This will be less than the pair's number of common relations or the number of common authors with whom both authors in the pair have published, as this "edge degree" does not include the strictly pairwise relationships. Moreover, this degree may help to characterize or model the group structures in the network.

Given the definition of vertex degree for graphs, it seems natural to analogously define the *degree for a k-simplex* as

$$d^{(k)}(\sigma_i^{(k)}) = L_{i,i}^{(k)}(\Delta). \tag{17}$$

This is indeed the usual definition.[18] However, this is not the number of $(k+1)$-simplices incident to $\sigma_i^{(k)}$, but rather Eq. (17) is the sum of the upper and lower degree of the k-simplex, i.e.,

$$\begin{aligned} \hat{d}^{(k)}(\sigma_i^{(k)}) &= \text{upper degree of } \sigma_i^{(k)} = (\boldsymbol{B}^{(k+1)}\boldsymbol{B}^{(k+1)T})_{i,i} \\ \check{d}^{(k)}(\sigma_i^{(k)}) &= \text{lower degree of } \sigma_i^{(k)} = (\boldsymbol{B}^{(k)T}\boldsymbol{B}^{(k)})_{i,i}, \end{aligned} \tag{18}$$

respectively. The *upper degree of $\sigma_i^{(k)}$* is the number of $(k+1)$-simplices incident to $\sigma_i^{(k)}$, whereas the *lower degree of $\sigma_i^{(k)}$* is the number of $(k-1)$-simplices incident to $\sigma_i^{(k)}$. Note that the lower and upper degree of k-simplices are determined by the diagonal elements of the first and second term of the kth combinatorial Laplacian in Eq. (10). Of course, the lower degree is fixed by definition to be number of boundary faces of a simplex, e.g., there are always 3 edges in every triangle. Hence, the lower degree of any k-simplex is always $\check{d}^{(k)}(\sigma^{(k)}) = \binom{k+1}{k} = k+1$.

For the collaborations in the C&N CTA dataset, the top-ranked pairs and triples of authors are given in Tables 12 and 13. It is difficult to compare these lists with the lists in Table 9 since the number of vertices (authors) is different than the number of edges (pairs of authors) and both are different than the number of 2-simplices (trios of authors). We can determine among the author pairs and trios with top simplex degrees how many of them also are authors with high centrality measures. For example, of the 27 author pairs listed in Table 12, only in 5 cases are both authors not included in the list of authors with high (graph) degree in Table 9, whereas this absence occurs in 23 and 19 cases for authors with high closeness and betweenness, respectively. Clearly, Table 13 is dominated by permutations of some subset of authors. This is due to a paper with 10 co-authors, where every trio of these authors starts with a degree of at least 10 (7 other authors plus 3 edges in their triangle).

Table 12 C&N CTA: Edge degree rankings

Rank	Author 1	Author 2	$d^{(1)}(\cdot)$
1	Myung Jong Lee	Tarek N Saadawi	19
2	Mariusz A Fecko	Sunil Samtani	18
3	Maitreya Natu	Adarshpal S Sethi	17
3	Anthony J McAuley	Raquel Morera	17
5	Richard Gopaul	Daniel Sterne	16
6	Kyriakos Manousakis	Anthony J McAuley	15
6	Geoff Lawler	Daniel Sterne	15
8	Sunil Samtani	M Umit Uyar	14
8	Natalie Ivanic	Daniel Sterne	14
8	Natalie Ivanic	Geoff Lawler	14
8	Georgios B Giannakis	Xiaoli Ma	14
8	Mariusz A Fecko	M Umit Uyar	14
13	Peter Kruus	Daniel Sterne	13
13	Richard Gopaul	Peter Kruus	13
13	Peter Budulas	Daniel Sterne	13
13	Peter Budulas	Richard Gopaul	13
17	Georgios B Giannakis	Shengli Zhou	12
17	Ahmed Abd El Al	Mariusz A Fecko	12
19	Ananthram Swami	Lang Tong	11
19	Brian Rivera	Daniel Sterne	11
19	Geoff Lawler	Brian Rivera	11
19	Peter Kruus	Brian Rivera	11
19	Peter Kruss	Geoff Lawler	11
19	John E Kleider	Xiaoli Ma	11
19	Richard Gopaul	Brian Rivera	11
19	Richard Gopaul	Geoff Lawler	11
19	Michalis Faloutsos	Srikanth V Krishnamurthy	11

Table 13 C&N CTA: Triangle degree rankings

Rank	Author 1	Author 2	Author 3	$\boldsymbol{d}^{(2)}(\cdot)$
1	Natalie Ivanic	Geoff Lawler	Daniel Sterne	14
1	Mariusz A Fecko	Sunil Samtani	M Umit Uyar	14
3	Richard Gopaul	Peter Kruus	Daniel Sterne	13
3	Peter Budulus	Richard Gopaul	Daniel Sterne	13
5	Geoff Lawler	Brian Rivera	Daniel Sterne	11
5	Peter Kruus	Brian Rivera	Daniel Sterne	11
5	Peter Kruus	Geoff Lawler	Daniel Sterne	11
5	Peter Kruus	Geoff Lawler	Brian Rivera	11
5	Richard Gopaul	Brian Rivera	Daniel Sterne	11
5	Richard Gopaul	Geoff Lawler	Daniel Sterne	11
5	Richard Gopaul	Geoff Lawler	Brian Rivera	11
5	Richard Gopaul	Peter Kruus	Brian Rivera	11
5	Richard Gopaul	Peter Kruus	Geoff Lawler	11

We can make comparisons between metrics of group centrality in graphs[23] with the simplex degrees in complexes. These group centrality extensions are given by

$$\begin{aligned} C_d(e_i) &= \sum_{j\neq i} |\boldsymbol{B}^{(1)}_{i:}\boldsymbol{B}^{(1)}_{:j}|, \\ C_c(e_i) &= \frac{n-2}{\sum\limits_{v_j\notin e_i} d(v_j,e_i)}, \qquad \text{and} \\ C_b(e_i) &= \frac{\sum\limits_{j<k} g_{jk}(e_i)/g_{jk}}{(n-2)(n-3)/2}, \end{aligned} \tag{19}$$

where $d(v_j, e_i)$ is the length of the shortest path from v_j to either vertex composing edge e_i, g_{jk} is the number of geodesics between vertices v_j and v_k, and $g_{jk}(e_i)$ is the number of geodesics between v_j and v_k that contain a vertex of e_i. Here, we only consider the (graph) group centrality metrics for adjacent vertices, although the metrics are more general. The (Pearson) correlation of the (simplex) edge degree with these centrality metrics (in the primary component) is

$$\text{corr}\begin{pmatrix} d^{(2)}(\cdot) \\ C_d(\cdot) \\ C_c(\cdot) \\ C_b(\cdot) \end{pmatrix} = \begin{bmatrix} 1.0000 & 0.1609 & -0.1653 & -0.1038 \\ 0.1609 & 1.0000 & 0.6185 & 0.7893 \\ -0.1653 & 0.6185 & 1.0000 & 0.7479 \\ -0.1038 & 0.7893 & 0.7479 & 1.0000 \end{bmatrix}. \tag{20}$$

It is clear that the edge upper degree is something different from the (edge) group centrality measures for graphs, which are in general more correlated than the individual centrality measures for graphs.

This notion of simplex degree is easily extendable. For example, the group structure of a network might be partly characterized by the number of triangles connected to a vertex, or the vertex-to-triangle degree. This can be found by adding the number of edge-to-triangle degrees for each edge connected to the vertex and then dividing by 2 (since each triangle connected to a vertex must be connected to 2 edges connected to that vertex and would be counted twice). More formally, this can be expressed as

$$\begin{aligned}\text{vertex-to-triangle degree}(v_i) &= \text{abs}(\boldsymbol{B}^{(1)}_{i,:})\text{abs}(\boldsymbol{B}^{(2)})/2 \\ &= \sum_k \sum_j \left| B^{(1)}_{i,j} \right| \left| B^{(2)}_{j,k} \right| /2. \end{aligned} \tag{21}$$

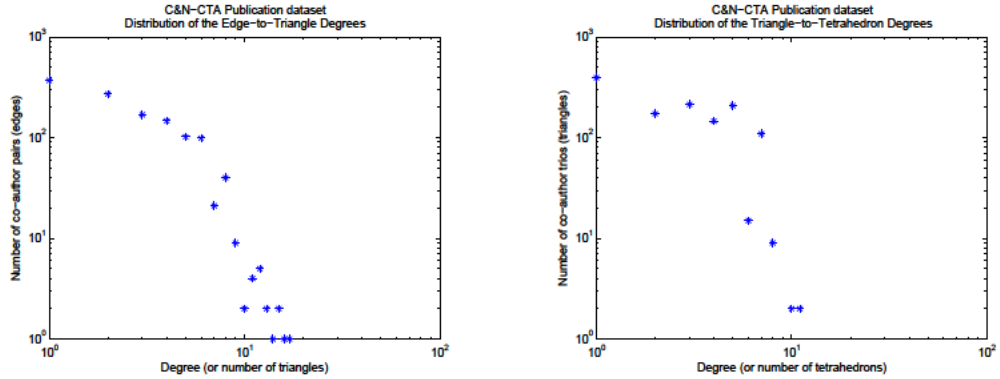
By an inductive argument, the vertex-to-k-simplex degree can be found by taking the summation over each dimension of the number of $(j-1)$-simplex-to-j-simplex degrees for each j-simplex adjacent to the vertex and dividing this total by $k!$, or simply

$$\sum_{j_1, j_2, \ldots, j_k} \left| \boldsymbol{B}^{(1)}_{i,j_1} \right| \left| \boldsymbol{B}^{(2)}_{j_1,j_2} \right| \cdots \left| \boldsymbol{B}^{(k)}_{j_{k-1},j_k} \right| /k!. \tag{22}$$

To generalize this extension beyond the vertex, the m-simplex-to-k-simplex degree is inductively calculated as

$$\begin{aligned}&m\text{-simplex-to-}(m+k)\text{-simplex degree}(\sigma^{(k)}_i) \\ &= \text{abs}(\boldsymbol{B}^{(m+1)}_{i,:})\text{abs}(\boldsymbol{B}^{(m+2)}) \cdots \text{abs}(\boldsymbol{B}^{(k)})/(k-m)! \\ &= \sum_{j_1, j_2, \ldots, j_k} |B^{(m+1)}_{i,j_1}||B^{(m+2)}_{j_1,j_2}| \cdots |B^{(m+k)}_{j_{k-1},j_k}|/k! \end{aligned} \tag{23}$$

For the C&N CTA co-authorship network, Figs. 13, 14, and 15 display various distributions of these simplicial complex degrees beyond the graph. It should be noted that for the m-simplex-to-k-simplex degrees, since these are log-log plots, the m-simplices having either 0 or 1 are counted together for $m < k-1$ or $m > 1$. At least for this dataset, there do not seem to be clear models for the m-simplex-to-k-simplex degree distributions when $m < k-1$. However, when $m = k-1$, the data possibly exhibit power law-like behavior over certain ranges.



(a) Edge Degree

(b) Triangle Degree

Fig. 13 The C&N CTA co-authorship network: Simplex degree distributions

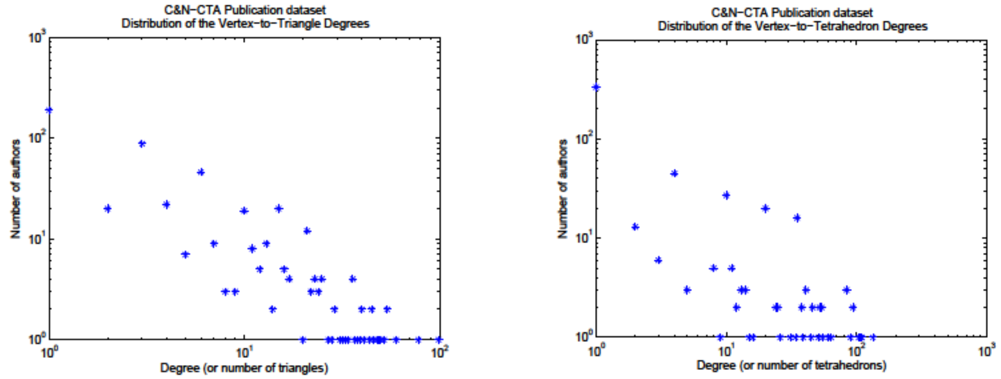


Fig. 14 The C&N CTA co-authorship network: Vertex-to-triangle and vertex-to-tetrahedron degree distributions

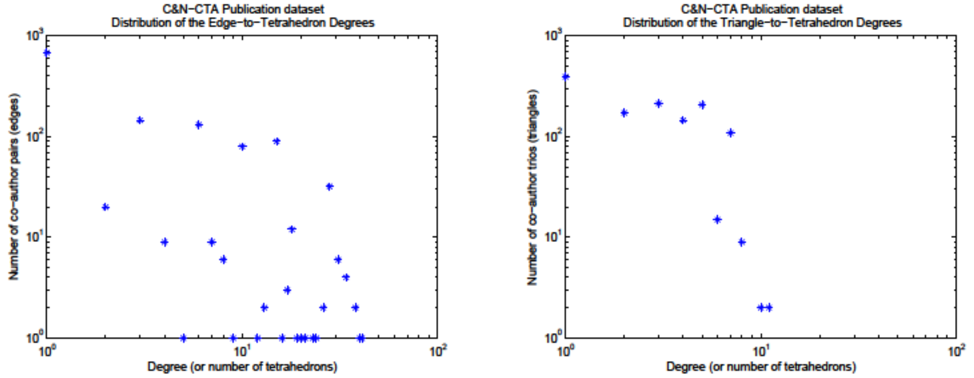


Fig. 15 The C&N CTA co-authorship network: Edge-to-tetrahedron and triangle-to-tetrahedron degree distributions

3.4.1.1 Facet Degrees

While characteristics of the upper degrees and the upper-degree distribution may certainly be important in modeling simplicial complex networks, the utility of higher-order simplex degrees is less clear. Perhaps a more interesting alternative extension to the m-simplex-to-$(m+k)$-simplex degree (for $k > 1$) is the consideration of the number of facets incident to a vertex (or possibly a simplex). In a collaboration network context, each facet represents the number of different groups within which a social actor interacts. Hence, the *facet degree* of an author is the number of distinct (maximal) collaborative groups in collaboration with the author.

The calculation of a vertex's facet degree involves finding how many k-simplices incident to the vertex are not incident to any $(k+1)$-simplices incident to the vertex for each k. More formally,

$$\text{facet degree}(v) = \sum_{k \geq 1} \text{card} \left\{ \sigma^{(k)} \ni v : \sigma^{(k)} \nsubseteq \sigma^{(k+1)} \ni v \right\}. \tag{24}$$

Similarly, the calculation of the facet degree of a simplex involves determining how many k-simplices are incident to the simplex but not incident to any $(k+1)$-simplex for each k, i.e.,

$$\text{facet degree}(\sigma) = \\ \sum_{k} \text{card} \left\{ \tau^{(k)} \supset \sigma : \nexists \tau^{(k+1)} \text{ s.t. } \tau^{(k)} \subset \tau^{(k+1)} \right\}. \tag{25}$$

Given a facet list, this can be computed with a relatively straightforward searching and counting procedure. The distribution of the facet degrees of the vertices in the C&N CTA network is shown in Fig. 16. An estimate of the exponent in the power-law relation is given by

$$\text{\#vertices} \propto \text{\#degree}^{-1.97}, \tag{26}$$

using a simple linear regression on the logarithm of the non-extremal values.

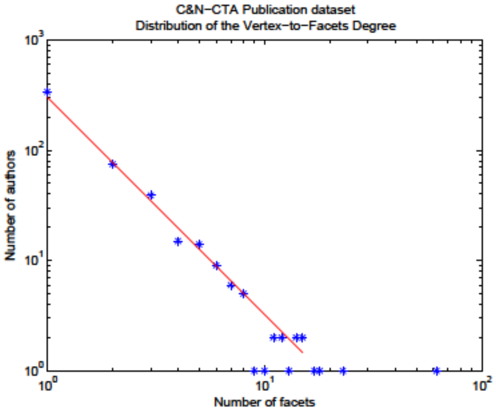


Fig. 16 The C&N CTA co-authorship network: Vertex-to-facet degree distribution.

3.4.2 Homology

Topological spaces can be classified by the number of holes in each dimension. The holes are not found in the topology but are defined by the number of homologous-independent cycles that do not bound. For each such cycle, all homologous cycles form an equivalence class that is a generator in the homology group. Moreover, for a given simplicial complex, we can characterize homology further by determining a "location" for each hole, or rather for each hole's boundary, defined as a cycle with minimal path length in the class. This is not a traditional abstract topological concept because homeomorphisms can change the geometric realization and, hence, where a particular location is perceived to reside. However, this approach has proven fruitful in determining gaps in coverage for sensor networks.[18,24]

For a collaboration network, holes in the simplicial complex may prove useful for characterizations and modeling of the network. The number of holes in each dimension potentially could significantly vary from what might typically be found in the sensor network coverage problems. This is because the sensor topology is ultimately characterized by the nature of its physical geometry, whereas a collaboration is not directly bounded by dimensions of physical space. As such, we can expect hole locations to be incident to k-simplices of relatively high degree and high clustering with $(k+1)$-simplices with relatively low degree.

The small-world nature of large co-authorship implies that over time authors will form relationships with other authors that are at some distance away in the network. Each instance of such a collaboration between 2 authors generally creates a new cycle that does not bound unless the relationship includes all authors on some path or chain between the 2 authors.

Figure 17 shows the homology for the cumulative network of the C&N CTA over the timeline of the program. The number of connected components has been discussed earlier. The number of holes is monotonically increasing. In fact, no hole when formed is ever "filled in" by latter collaborations in this dataset. This is counterintuitive to the notion of the small-world properties in co-authorship networks. Because of the presumed high clustering, connected triples on cycles defining hole locations should shrink over time so long as the authors remain active. This property should be present in larger networks observed over a sufficient timeline. However, it is likely the number of newly generated holes will still outpace the holes that get filled in.

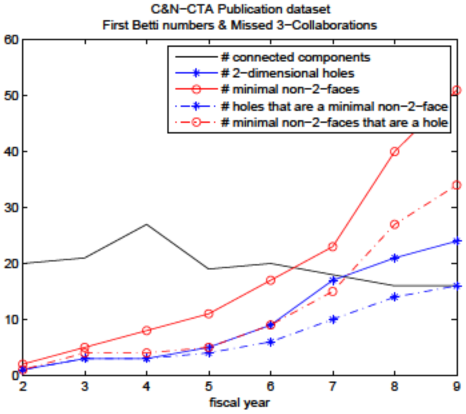


Fig. 17 Comparison of number of components, holes, and minimal non-faces over the evolution of the network

An interesting feature of this particular network is that all holes are in the primary component (all other components are tree-like complexes). In addition, the subcomplex induced by the inclusion map restricted to the vertices adjacent to the hole locations, i.e., the vertices comprising the shortest cycles that do not bound, form a single component. As seen in Figs. 17 and 7, most of the holes are small (a hole located by a 3-edge cycle is a minimal non-face, see below); however, there is a hole with shortest bounding cycle of 14 edges (or vertices) linking many of the smaller holes together. An interesting question for future work is whether this feature is inherent in much larger networks.

3.4.3 Minimal Non-Faces

A *minimal non-face* is a set of vertices in a simplicial complex Δ such that every subset except the set itself is a simplex in the complex, and we call it a *minimal non-k-face* if the set of vertices has dimension k. A minimal non-k-face in a collaboration network may indicate the potential for a larger collaborative connection. Although a minimal non-k-face is a hole in the subcomplex comprising only the simplices of the vertices in the missing face, it is not necessarily a hole in the entire complex. For example, if each pair of vertices in a minimal non-2-face is collaborating with another common author (geometrically, this is an empty and bottomless tetrahedron), then there is no hole.

Similar to the case with holes, minimal non-faces should be incident to simplices with high degree and centrality. The distinction is that holes may form from connections made between distance parts of the network, whereas minimal non-faces are formed explicitly by independent local clustering. For example, a triangle in the graph representation can either exist because of a 2-simplex relation or a minimal non-face relation in the simplicial complex.

Figure 17 shows the growth of the number of minimal non-faces for the cumulative network of the C&N CTA, as well as a comparison between holes and minimal non-faces. Not every minimal non-2-face corresponds to a hole nor does every hole correspond to a minimal non-2-face. Clearly, minimal non-faces are more prevalent than holes. It is possible for the number of minimal non-faces corresponding to holes to be greater than the number of holes because of the non-uniqueness of shortest cycles identifying a hole location.

3.5 Miscellaneous

3.5.1 Q-analysis

In this section, we consider labeled simplices, which are analogous to hyperedges. Not every simplex has a label. Only those corresponding to a relation have a label, and it is possible for multiple labels to exist for a simplex. When we refer to a simplex, we are referring to a labeled simplex. In the co-authorship network context, each labeled simplex corresponds one-to-one to a paper in the dataset.

Two simplices are *q-connected* if there exists a finite sequence of q-simplices such that each consecutive pair of simplices share a q-dimensional face, i.e., each consecutive pair is *q-near*. The length of this sequence, or *chain of q-connection*, is 1 less than the number of simplices in the sequence. Hence a q-simplex is q-connected to itself by a chain of length zero.

The $\check{q}$ number of a simplex is the greatest q for which the simplex is q-connected to a distinct simplex. (If the simplex is a face of another simplex, then $\check{q}$ is the dimension of the simplex.) The $\hat{q}$ number of a simplex is simply the dimension of the simplex. The *eccentricity* of a simplex is defined as

$$\mathrm{ecc}(\sigma^{(q)}) = \frac{\hat{q} - \check{q}}{\check{q} + 1} = \frac{q - \check{q}}{\check{q} + 1}, \tag{27}$$

and is a measure of the individuality of a simplex. Since eccentricity is undefined when $\check{q} = -1$ (i.e., when the simplex is isolated), it is convenient to use normalized eccentricity,[26] defined as

$$\mathrm{ecc}(\sigma^{(q)}) = \frac{\hat{q} - \check{q}}{\hat{q} + 1}. \tag{28}$$

With this normalization, an isolated simplex has eccentricity 1 and a simplex that is a subset of another has eccentricity zero.

As q-connectivity is a relation, the simplices with dimension at least q can be partitioned into equivalence classes in which 2 simplices are in the same equivalence class or q-connected component if they are q-connected. The number of such components in the simplicial complex for a given q is denoted by Q_q, and the vector

$$\boldsymbol{Q} = \begin{bmatrix} Q_0 & Q_1 & \dots & Q_m \end{bmatrix}^T \tag{29}$$

is called the *first structure vector* of the complex. (Note that the simplices with dimension at least q are the simplices left over after deleting the simplices/sets from the $(q-1)$-skeleton from the simplicial complex. Collectively, these are not a simplicial complex.)

For the final year of the cumulative network of the C&N CTA dataset, the first structure vector is

$$\boldsymbol{Q} = \begin{bmatrix} 16 & 139 & 240 & 99 & 37 & 21 & 9 & 5 & 1 & 1 \end{bmatrix}^T. \tag{30}$$

Comparing this with the number of labeled simplices that exist for each q, i.e., $(960, 937, 523, 151, 48, 22, 10, 6, 2, 1)$, we see that the labeled simplices are very weakly connected at dimension $k > 2$.

3.5.2 Strong Collapsing

Every simplicial complex has a conjugate complex in which the labeled simplices are mapped to vertices and simplices exist in the conjugate among vertices whose labeled simplices have non-empty intersection. This conjugate complex is similar to the nerve of a complex. It can be shown that the simplicial complex and conjugate complex have the same homology.[27]

Note that when the eccentricity of a simplex is zero, then the simplicial complex is essentially the same without that simplex. Any subset of the simplex is covered by some other simplex. Since the complex is unchanged, then homology is preserved even if that simplex is removed. Hence the homology is preserved in the conjugate when the vertex corresponding to the labeled simplex with nil eccentricity is removed (or collapsed). This motivates an iterative process of collapsing vertices in the conjugate complex and the original complex to reduce the dimension and size of the original complex while maintaining its homology. This is called *strong collapsing*.[28]

When applied to the cumulative network of the C&N CTA, the network collapses to a core 67 vertices (authors) with 78 labeled simplices (papers). The non-primary components collapse to single vertices. The primary component collapses to a connected set of vertices that include all the authors with high centrality metrics. Moreover, strong collapsing preserves hole locations in addition to preserving the holes, i.e., at least 1 shortest cycle incident to the hole remains after strong collapsing.

3.5.3 f-vector and Euler Characteristic

The f-vector (or face vector) of a simplicial complex is simply a vector whose ith element indicates the number of $(i-1)$-simplices. For the C&N CTA, this is

$$\boldsymbol{f}^T = \begin{bmatrix} 518 & 1248 & 1272 & 952 & 648 & 362 & 156 & 49 & 10 & 1 \end{bmatrix}. \tag{31}$$

This can also be compared with the first structure vector; however, this creates the appearance of a greater number of potential components than exist since many of these simplices are from the large facets. As mentioned earlier, 252 of the 648 4-simplices are due to a single facet (paper).

The Euler characteristic χ is a topological invariant, which might be useful in distinguishing different network types. It can be calculated by either an alternating sum of the elements of the f-vector or an alternating sum of the Betti numbers, i.e.,

$$\begin{aligned}\chi &= f_1 - f_2 + f_3 - f_4 + \cdots &(32)\\ &= b_0 - b_1 + b_2 - b_3 + \cdots &(33)\end{aligned}$$

where f_k represents the number of $(k-1)$-dimensional simplices in the complex and b_k represents the kth Betti number.

For the C&N CTA yearly and cumulative network, Table 14 displays the Euler characteristics. It is not surprising that since the number of holes is increasing whereas the number of components remains relatively the same (see Fig. 17), the Euler characteristic gradually decreases over time for the cumulative network. For the yearly network, the characteristic remains positive and oscillates near 20 because the holes are created over multiple years rather than in a single year.

Table 14 C&N CTA: Euler characteristic

Fiscal Year	02	03	04	05	06	07	08	09
Euler char. (cum.)	19	18	24	14	11	1	-5	-8
Euler char. (per year)	19	13	31	21	24	14	16	17

3.5.4 Visualization

The visualization of a simplicial complex is more difficult than that of a graph due to the need to represent the groupings instead of just the edges among the vertices. Theoretically, a k-dimensional simplicial complex can be geometrically realized in $2k+1$ dimensions.[16,29] This is a difficult task for what ultimately is to be projected onto a 2-dimensional picture of 3-dimensional space. Therefore, to represent the different size of the groups we use color.

An investigation of the existing visualization schemes for simplicial complexes revealed Polymake[30] as the best for small simplicial complexes. Unfortunately, Polymake did not distinguish between different orders of simplices well because it did not have an option to use color based on size. Another shortcoming of Polymake is the inability to visualize disconnected components simultaneously. Furthermore, Polymake's visualization algorithm did not have clear options to modify its routine. Thus, for a simplicial complex with a larger number of vertices, the spring-embedding component of Polymake's algorithm did not converge and no visual output is produced. Simply put, the simplicial complex created from the C&N CTA co-authorship dataset proved too large. Primarily to cope with this last failure, we created code for visualization within MATLAB.

A visualization is of the C&N CTA is shown in Fig. 7. A greedy algorithm is used to place the vertices. Then a spring embedder is applied on this initialization. The spring force encourages a $\binom{k+1}{3}$-sided polyhedron (imposed in a 2-dimensional realization of 3-dimensional space) for k-simplices while also creating separation between disconnected vertices and simplices.

This approach includes several desirable features that Polymake does not as listed above: color represents the simplex order, visualizations of disconnected components are possible, and the approach handles the number of vertices in the C&N CTA simplicial complex without convergence issues.

4. Conclusions

A higher-dimensional analysis of ARL's C&N CTA co-authorship network was presented. This network represents the scientific collaborations among researchers that resulted in publications. Although the network is medium-sized, many of the graph theoretical measures and metrics are similar to what is commonly found in large co-authorship or other social networks. This justifies the utility of this network in our study of higher-dimensional collaborative structures.

The higher-dimensional structures were modeled as simplices in a simplicial complex. The connectivity of the simplicial complex is determined by the facets and their intersections. Facets represent particular collaborative groups and their intersections represent the interactions between the groups. This makes properties related to the facets of the simplicial complex key to characterizing the collaborative structure of the these networks. We have demonstrated that the facet degree distribution follows a power law relationship. The homology of the simplicial complex can be characterized by the minimal length cycles that do not bound. These homology cycles have connections to short cuts in small world networks. In fact, such 1-cycles that are not non-minimal faces consist exclusively of short-cut edges. In a sense, these homology cycles can be viewed as a form of global clustering. We have shown that the cycles tend to intersect at authors with high centrality measures in the graph sense. We have also shown that minimal non-faces represent a form of independent local clustering, which is also an innate feature of small world networks.

5. References and Notes

1. ARL: Collaborative Alliances: Completed CTAs. Adelphi (MD): United States Army Research Laboratory; 2011 March [accessed 2014 March 1]. www.arl.army.mil/.

2. If it were known how many doctorates were attained by students supported by the program in each FY, then one might expect a larger than average number in FY05. Unfortunately, this information was not available at the time of this report.

3. Hsu JW, Huang DW. Distribution for the number of coauthors. Physical Review E. 2009;80:057101-1–4.

4. Newman MEJ. Coauthorship networks and patterns of scientific collaboration. Proceedings of The National Academy of Sciences. 2004;101:5200–5205.

5. Newman MEJ. Scientific collaboration networks. I. Network construction and fundamental results. Physical Review E. 2001;64:016131-1–8.

6. Newman MEJ. 2010. Networks: An Introduction. Oxford (United Kingdom): Oxford University Press.

7. Wasserman S, Faust K. 1994. Social Network Analysis: Methods and Applications. New York (NY): Cambridge University Press.

8. West DB. 2001. Introduction to Graph Theory. Upper Saddle River (NJ): Prentice Hall.

9. Technically, the author vertices are linked to paper vertices in a bipartite graph where the link or edge between an author and paper indicates that the author is 1 of the co-authors of the paper. The co-authorship graph, then, is the “one-way mode projection” of this bipartite graph onto the author vertices.

10. Liu X, Bollen J, Nelson ML, Van de Sompel H. Co-authorship networks in the digital library research community. Information Processing and Management. 2005;41:1462–1480.

11. Zheleva E, Sharara H, Getoor L. Co-evolution of social and affiliation networks. In: ACM. Proceedings of The 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; 2009 June 28–July 01; Paris (France). New York (NY): ACM; c2009. pp. 1007–1016.

12. For a $G(n, p)$ model, the degree has a Bin$(n-1, p)$ binomial distribution and approaches a Pois(np) Poisson distribution as n grows; hence, the ratio of the average degree to the graph size approaches $p = 0.009$.

13. Barabási AL, Jeong H, Néda Z, Ravasz E, Schubert A, Vicsek T. Evolution of the social network of scientific collaborations. Physica A. 2002;311:590–614.

14. Ramasco JJ, Dorogovtsev SN, Pastor-Satorras R. Self-organization of collaboration networks. Physical Review E. 2004;70(3):036106:1–10.

15. Hatcher A. 2002. Algebraic Topology. Cambridge (MA): Cambridge University Press.

16. Munkres JR. 1984. Elements of Algebraic Topology. Cambridge (MA): Perseus.

17. For $k = 1$, this should not be confused with the concept of paths in a graph. Whereas, each path corresponds to an element of the 1-chains, each path also has an ordering on the fundamental basis, the edges, making up that element. Moreover, an element of the 1-chains can have weights that do not correspond to any single path in the graph.

18. Muhammad A, Egerstedt M. Control using higher order Laplacians in network topologies. Paper presented at: MTNS 2006. Proceedings of the 17th International Symposium on Mathematical Theory of Networks and Systems; 2006 July 24–28; Kyoto (Japan). p. 1024–1038.

19. The characteristics for the active network are close to the yearly network and so are not shown in this analysis.

20. This would be an annotation to the simplicial complex and not a *weighted simplicial complex* as it is commonly defined.

21. We only detail the linear models for Fig. 10d since this is the case with the strongest evidence of a linear relationship.

22. Klivans CJ, Nyman KL, Tenner BE. Discrete Mathematics. 2009;309:4377–4383.

23. Everett MG, Borgatti SP. Extending centrality. In: Carrington PJ, Scott J, Wasserman S, editors. Models and Methods in Social Network Analysis. New York (NY): Cambridge University Press; 2005. p. 57–76.

24. de Silva V, Ghrist R. Coordinate-free coverage in sensor networks with controlled boundaries via homology. The International Journal of Robotics Research. 2006;25(12):1205–1222.

25. Not to be confused with eccentricity of a vertex in a graph.

26. Maletić S, Rajković M, Vasiljević D. Simplicial complexes of networks and their statistical properties. In: Bubak M, van Albada G, Dongarra J, Sloot PMA, editors. ICCS 2008. Proceedings of the International Conference on Computational Science, Lecture Notes in Computer Science 5102; 2008 June 23–25; Kraków (Poland). Berlin (Germany): Springer, c2008. p. 568–575.

27. Dowker CH. Homology groups of relations. Annals of Mathematics, 2nd Series. 1952;56(1):84–95.

28. Wilkerson AC, Moore TJ, Swami A. Simplifying the homology of networks via strong collapses. Paper presented at: ICASSP 2013. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing; 2013 May 26–31; Vancouver (Canada). p. 5258–5262.

29. van Kampen ER. Komplexe in euklidischen Räumen. Abhandlungen aus dem Mathematischen Seminar der Universität Hamburg. 1933;9(1):72–78.

30. Gawrilow E, Joswig M. Polymake: a framework for analyzing convex polytopes. In: Kalai G, Ziegler GM, editors. Polytopes—Combinatorics and Computation. Basel (Switzerland): Birkhäuser; 2000. p. 43–74.

1	DEFENSE TECHNICAL
(PDF)	INFORMATION CTR
	DTIC OCA
2	DIRECTOR
(PDF)	US ARMY RESEARCH LAB
	RDRL CIO LL
	IMAL HRA MAIL & RECORDS MGMT
1	GOVT PRINTG OFC
(PDF)	A MALHOTRA
4	US ARMY RSRCH LAB
(PDF)	RDRL CIN T
	TERRENCE J MOORE
	ROBERT J DROST
	ANANTHRAM SWAMI
	BRIAN M RIVERA
1	US ARMY RSRCH LAB
(PDF)	RDRL CIN
	ALEXANDER KOTT